AD_____

Award Number: DAMD17-00-1-0410

TITLE: Remote Patient Management in a Mammographic Screening Environment in Underserved Areas

PRINCIPAL INVESTIGATOR: David Gur, Sc.D

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, PA 15260

REPORT DATE: September 2004

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20050113 048

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE September 2004	3. REPORT TYPE AND DATES COVERED Annual (1 Sep 2003 – 31 Aug 2004)

4. TITLE AND SUBTITLE
Remote Patient Management in a Mammographic Screening Environment in Underserved Areas

5. FUNDING NUMBERS
DAMD17-00-1-0410

6. AUTHOR(S)

David Gur, Sc.D

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pittsburgh
Pittsburgh, PA 15260

E-Mail: gurd@msx.upmc.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

Early detection of breast cancer is of significant interest to our society. Mammographic screening is gradually moving toward a "distributed acquisition – centralized review" approach. Unfortunately, a relatively high recall rate using this approach increases patient anxiety as well as the cost and complexity of the diagnostic process. The purpose of this project is to evaluate in a multi-phase project the possible impact of a unique tele-mammography system that utilizes common carriers with wavelet-based data compression for image transmission, on the recall rate in remote locations where physicians are not available during mammographic procedures. The initial phases of the project encompasses the design, assembly, and technical testing of a multi-site tele-mammography system that enables the digitization, transmission, and display of wavelet compressed images, as well as associated text documents of a case combined with CAD results in less than 15 minutes. The possible impact of such a system is being evaluated during a step-by-step assessment in a multi-site study.

14. SUBJECT TERMS
Breast Cancer, Telemammography, Detection, CAD

15. NUMBER OF PAGES
54

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction and Background:

Please note that this section of the report remains largely the same as reported in previous years. Periodic mass screening of asymptomatic women is rapidly gaining approval and acceptance, and the population segment recommended for screening is increasing due to both longer life expectancy as well as earlier recommended age for initial examination [1-3]. The large variability in a number of important aspects related to mammography, as practiced in the U.S., resulted in the enactment of the Mammography Quality Standards Act, which mandates accreditation of each program (facility, technical, and professional) [4,5]. Shortages of expert mammographers in many locations, combined with the desire to make it convenient for the patient to undergo the procedure, suggest that there may be a need for high-quality tele-mammography systems that enable a distributed acquisition-centralized expert review type solution to the problem, particularly in underserved areas [6, 7]. The relatively high recall rates (5-15%) of screened women to supplement information that was not ascertained during the initial visit (e.g. magnification views, ultrasound) also make it desirable to enable physician "monitoring" and "management" of remote underserved locations so that some patient-management decisions can be made while the patient remains in the clinic [8-11]. In addition, a technologist who observes a possible abnormality during the performance of the study could benefit from the ability to communicate her/his suspicion, and an expert mammographer could review the specific case, together with the technologist's observation, resulting in an improved and perhaps a more timely diagnosis. Current practices result in increased patient anxiety and added practice complexity and cost. Early attempts to develop and implement a practical tele-mammography solution to this problem failed due to several significant technical problems associated with acquisition, transmission, management, and display of the images and other related information [12-14]. Many of these technical issues have been resolved in recent years, but some remain [14-18]. Although an adequate communication infrastructure for high-quality tele-mammography is available within some urban regions, the fact remains that where it may be needed most (i.e. remote, non-urban locations), enabling (two-way) communication systems remain limited to lower level communication capabilities. Other communication technologies, such as satellites, are being evaluated for this purpose, but it is not likely that these will displace lower level communication technologies in many underserved areas for quite some time [19-23]. Hence, the problem of cost effective, timely remote patient monitoring and management in many underserved areas is not a simple one.

As a part of this project, we are assembling and evaluating a unique tele-mammography system that enables improved communication between remote sites where physicians are not always available during the mammographic acquisition process and a central location where experts can review the acquired images shortly after acquisition and assess whether or not additional procedures (e.g., spot compression views) are needed [24, 25]. The system we are assembling and testing is based on prior preliminary experience acquired in our group during ten years of research in this general area. It includes the use of a common carrier for communication (Plain Old Telephone System, POTS) and other "low level" communication capabilities, wavelet-based image compression for data reduction, and the optional incorporation of other text information, location information, and CAD results into the transmitted information. The main goal is to assess in a step-by-step approach whether the use of such a system could substantially reduce recall rates in the remote sites. Other objectives regarding ways to improve communication between the technologist at the

remote site and a radiologist at the central site, as well as creating an environment for "more active" participation of the technologist in the diagnostic process, are also being explored.

Body:

Since the initiation of the project on September 1, 2000, we have been progressing methodologically step by step on the tasks listed in the Statement of Work (page 5 of the proposal), as originally submitted. As will be explained in the body of this annual report, our initial findings resulted in the addition of several technical tasks that were successfully performed in order to maximize our ability to learn about the applications being investigated in this project. During year four of the project, work was performed in three different areas listed under Task 1 (Redesign and Assemble System), Task 3 (Clinical System's Evaluation), and Task 4 (Evaluation of CAD Results) in the original proposal. We have also begun planning for Task 5. As we explain in the body of the report, a significant new addition (capability) was added to the system as a result of our operational and preliminary clinically simulated evaluation tasks. This required a substantial technical effort and ultimately resulted in a major software upgrade of the system. The task was recently completed and the system has been tested using the new software. We recently requested a one year no-cost extension to the project to complete all tasks, including Task 5 (High Volume Demonstration). Following the recommendations of the reviewer of our latest annual report, we focus here on work performed during the period in question (September 1, 2003 to August 31, 2004).

Under Task 1, we performed the following:

Since our last progress report, and based on the results from our observer performance studies, we decided to incorporate into the system in an integrated, easy to use fashion: 1) text messaging (namely, two way "chat" between the remote technologist and central radiologists), 2) marking of suspicious locations (namely, the technologist marks suspicious regions on an image overlay), 3) CAD results, and 4) prior mammography reports. The reason for the additional tools (in particular item #2) is to provide the radiologist at the central location with all of the tools to enable better assessment of the examinations being sent for review (obviously, this is all done in addition to the actual mammographic images). Therefore, these technical tasks were planned for and implemented before we performed a "high-volume" simulated study (Task #5 in our original proposal). The driving force behind this additional work was the result of clear indications from our radiologists during a prior observer performance study that any additional information we can provide during the remote review of examinations would be of great help. Since the technologists at the remote sites send examinations that they already believe would require a follow-up procedure, it was felt that it would be important to communicate the specific (known) location of the "suspected" region on the images which the technologists identified as the reason for sending the case for a central review. The interfacing to enable these capabilities required substantial modifications of our software (in particular item #2), and we decided to perform an additional retrospective observer performance study to assess the ability to integrate this information in an easy to use manner, and evaluate potential (projected) reduction in recall rates when we add this function to the communication capabilities of the system.

We completed the development and technical testing of the new software (a major upgrade was installed and tested during the third week of May, 2004). Technical testing of the system was completed and an observer performance study was carried out (see Task 3).

Under Task 3, we performed the following:

1) Retrospective observer performance studies to assess performance without and with CAD results and without and with the location of suspected regions:

After the software upgrade was completed and tested (Task #1), a study management software routine for the retrospective reading experiment was written and tested during the third and fourth weeks of May, 2004. All data entry for a retrospective observer performance study (including selecting cases and related information) was completed in early June, 2004.

A synopsis of the three observer performance studies in this area follows: Registered mammography technologists from three remote imaging sites transmitted 245 screening mammography exams to a central site (radiologists), which they (the technologists) believed needed additional procedures. Four data components are transmitted from the remote site: (1) image data - current exam mammography films digitized at 50 μm pixel dimensions; (2) text and graphic communication between the technologist and the radiologist via a "chat" box in which the technologist can describe and mark suspicious regions on integrated generic images; (3) prior patient reports when available; and (4) computer aided detection (CAD) results. At the central site images are displayed on a workstation consisting of three high-resolution, portrait monitors. The image data with the CAD results overlaid are displayed on two monitors and the chat box and prior reports on the third monitor. Seven radiologists reviewed and rated the exams on the tele-mammography workstation and indicated: (1) if additional procedures were recommended, (2) when appropriate, which breast was involved, and (3) when appropriate, the specific recommended procedures. The performance of the radiologists on the workstation was compared with the clinical interpretation of the same examinations in three studies. Study 1 had two interpretation modes: (1) images only and (2) images and technologist's text message. Study 2 had two modes: (1) images and technologist's text message and (2) images, text message, and prior report. Study 3 had three modes: (1) images, technologist's text message, and prior report; (2) images, text message, prior report, and technologist's graphic location marks; and (3) images, text message, prior report, graphic marks (location), and CAD results. We are currently in the process of completing the analyses of the third study. Amongst other analyses, we will compute the potential improvements in terms of projected reduction in recall rates at the remote sites.

Preliminary Results: Technologists were able to identify suspicious examinations that may require additional procedures, but their "recommended" examinations amounted to a substantially larger number compared with that of a clinical interpretation by a radiologist. The 245 screening exams were successfully transmitted, processed, reviewed, and rated. The percent of exams recalled for recommended additional procedures (termed "recall") during the actual clinical interpretation for Studies 1 (n = 130), 2 (n = 99, a subset of Study 1), and 3 (n = 115) were 39.2%, 38.4%, and 42.2%, respectively. Tele-mammography Study 1; modes 1 and 2 had mean recall rates of 73.3% (+/- 17.9) and 82.5% (+/- 16.2), respectively, and mean agreements of 51.7% (+/- 5.5) and 48.7% (+/- 6.3), respectively. Study 2; modes 1 and 2 had mean recall rates of 79.6% (+/- 12.3) and 77.5% (+/- 13.8), respectively, and mean

agreements of 52.3% (+/- 6.7) and 52.8% (+/- 7.0), respectively. Study 3; modes 1, 2 and 3 had mean recall rates of 72.3% (+/- 9.3), 72.3% (+/- 9.3), and 72.7% (+/- 9.2), respectively, and mean agreements of 57.4% (+/- 4.6), 57.1% (+/- 3.9), and 56.7% (+/- 3.9), respectively. However, it should be remembered that without radiologists' reviews 100 percent of these women would have been recommended for additional procedures by the technologists; hence, approximately 30 percent reduction could be achieved utilizing our proposed approach. These results are preliminary and we hope to complete our analyses before September 30, 2004.

2) Clinical assessment of traditional performance levels:

A substantial fraction of the effort during the last year was carried out under Task 3. We are "breaking ground" in several respects that include but are not limited to the involvement of technologists in the decision-making process (namely, which cases to send over to the central site and why?), and possibly the increased "reliance" of the radiologists on the technologists' judgments. As a part of this investigation we assessed our clinical performance levels in the traditional practice (without tele-mammography).

We analyzed data available in our databases concerning patient distributions and process-related information. This includes the recall rate by physician, site, type, and reason for recall. We also reviewed records concerning the cycle time from the initial examination to a definitive diagnosis for cases that were not being recalled, as well as cases that were. This effort constitutes the reference information for comparison purposes. One of the more interesting (and relevant) findings in this regard is the long delays in scheduling (average > 20 days) between the patient's call for an appointment due to recall and the actual date of examination, underlying the potential benefit of the use of tele-mammography to reduce recall rates. During the last year, we completed a large study to assess the effect of the introduction of CAD into our clinical environment and the relationship between recall rates and detection rates for our ten highest volume radiologists. One of the issues that was raised in our group was the issue of correlations (if any) between the recall and detection rates of radiologists. This is an important point since there is a significant pressure on radiologists to reduce their individual recall rates to below ten percent. While we recognize the tremendous value of reducing recall rates without a substantial degradation in detection rates (sensitivity), the question arises as to whether or not higher recall rates are also generally associated with higher detection rates. These studies involved the reviews of over 115,000 records and resulted in important observations that were published in JNCI and Cancer (see publications list). We strongly believe that the use of CAD will ultimately be used as an integral part of the diagnostic process and some of our efforts to develop and improve CAD schemes were supported (only to a very minimal level) by this project. Several important observations were made, all of which were published (see publications list).

Under Task 4, Evaluations of CAD Results:

We continue to improve our own CAD schemes, and as we progress in this area we also change the performance level of the scheme used in the tele-mammography system. Most of the CAD development efforts are performed under separate projects, but limited effort related to the assessment of performance is relevant to this project (see publications list). As of a recent progress review and planning meeting (August 24, 2004), we determined

and fixed the operating point of the scheme (sensitivity and false positive rate) to be used in the system during the demonstration task (Task #5).

Under Task 5, Clinically simulated almost real time transmission and reporting:

There is only one significant effort under this category; namely, the performance of an "almost real time - high volume" demonstration of the transmission of suspected cases at the remote sites and a clinically simulated response from the central site. This task is planned for execution during our proposed no cost extension year, since we needed to upgrade the system and perform two additional observer performance studies. We already had two planning meetings regarding this task and we anticipate that once the management software for the study is completed it will be carried out in an "almost real time" simulated clinical environment. During this study, we anticipate that each site will transmit approximately 6-12 studies per day to the central site. We continue to test the system's and radiologists' ability to handle the workflow and review of this reasonably high volume of cases.

Key (Research) Accomplishments:

During the last year of the project, we have been progressing according to the original plan and addressed a large number of the technical tasks and operational issues associated with the design, implementation, technical, and clinically simulated testing of the multi-site tele-mammography system. The key accomplishments for the last year were:

- We carried out a comprehensive review of the performance of our radiologists in terms of performance without and with CAD, as well as the relationship between recall rates and detection rates.
- We upgraded the system with a major software revision in response to radiologists' preferences during the performance of the task the tele-mammography system was designed for.
- We successfully and reliably transmitted approximately 530 cases from three remote sites to the central site (our total to date exceeds 2000 cases).
- We successfully reviewed a large number of cases on the workstation and generated a clinically simulated response to the remote sites.
- We completed two observer performance studies to assess agreement levels between the technologists and radiologists on suspicious cases. The analyses of these studies are currently being performed and will be reported in the 2005 SPIE meeting.
- We are increasing the communication level between technologists and physicians in regard to decision-making processes, and we are engaged in discussions concerning a more extensive use of technologists as physician extenders in several areas.
- We demonstrated that in principle one can achieve a significant reduction in actual recall rates for a second visit, albeit at this time, at the cost of a substantial increase in the number of women who would receive additional procedures during their initial screening visit. We continue to focus on ways to reduce this number.

Reportable Outcomes:

The nature of this project is such that some of the technical work performed to date does not result in a large number of significant reportable outcomes. However, as we developed and tested the system, several reportable tasks have been performed. For example, our comprehensive assessment of the actual performance in our clinical operations as it relates to the use of CAD and to recall and detection rates was reported. These efforts have led to important developments and observations that may have a significant impact on this field. Therefore, several of our new publications which had not been reported in our last progress report, are listed below. Since our last report these include:

1. Leader JK, Sumkin JH, Ganott MA, Hakim C, Hardesty L, Shah R, Wallace L, Klym A, Drescher JM, Maitz GS, Gur D. Subjective assessment of high-level image compression of digitized mammograms. <u>Proceedings of SPIE Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment,</u> San Diego, CA, February 2004, 5372:415-422.
2. Zheng B, Hardesty LA, Poller WR, Sumkin JH, Golla S. Mammography with computer-aided detection: reproducibility assessment – initial experience, Radiology 2003; 228:58-62.
3. Zheng B, Good WF, Armfield DR, Cohen C, Hertzberg T, Sumkin JH, Gur D. Performance change of mammographic CAD schemes optimized with most-recent and prior image database, Acad Radiol 2003; 10:283-288.
4. Chang YH, Good WF, Leader JK, Wang XH, Zheng B, Hardesty LA, Hakim CM, Gur D, Integrated density of a lesion: a quantitative, mammographically derived, invariable measure, Med Phys 2003; 30:1805-1811.
5. Zheng B, Leader JK, Abrams G, Shindel B, Catullo V, Good WF, Gur D. Computer-aided detection schemes: The effect of limiting the number of cued regions in each case, AJR 2004; 182:579-583.
6. Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim CM, Hardesty L, Poller WR, Shah R, Wallace L. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system, JNCI 2004; 96:185-190.
7. Gur D, Sumkin JH, Hardesty LA, Clearfield RJ, Cohen CS, Ganott MA, Hakim CM, Harris KM, Poller WR, Shah R, Wallace LP, Rockette HE. Recall and detection rates in screening mammography. A review of clinical experience – implications for practice guidelines. Cancer 2004; 100(8):1590-1594.

Conclusions:

There are several technical, clinical, and assessment tasks listed in the Statement of Work of this project. During the first four years, we undertook a large number of technical and application-based tasks associated with the design, implementation, and preliminary evaluation of a multi-site tele-mammography system. We overcame many of the technical problems and assembled a multi-site system that exceeds several of the performance goals we originally proposed. The system has been undergoing a comprehensive step-by-step evaluation (and major upgrades as deemed appropriate). Our main observation to date is that the general concept was verified and the actual implementation resulted in an appreciation for the importance of the "comfort level" of the team (physicians and technologists) in

operating and using such a system for the stated purpose. Most important perhaps is the demonstration that in principle, one could achieve a significant reduction in actual recall rates for a second visit. At this time, it can be done at the cost of a substantial increase in the number of women who would receive additional procedures (e.g., views) during their initial screening visit, and we currently focus on investigating different ways to reduce this number. In addition, we have improved substantially our understanding of several extremely important issues related to screening mammography, in general, and the use of CAD, in particular. These may have far reaching implications on this field.

So What?

The main goal of this project is to evaluate how the use of an "almost real-time" tele-mammography system (with or without the use of CAD results and other relevant information) may impact the diagnostic process in terms of complete cycle time and patients' recall rate. Task 5 (a high volume clinically simulated study) is planned as the last major effort for this project. Success of this project will enable a comprehensive demonstration of different ways to increase communication between remote (and potentially underserved) sites and a central site. Our hope is that by using this approach, one may be able to provide better, more timely and cost-effective service at these sites, and in the process substantially reduce actual recall rates in these remote facilities. Despite significant advances in our understanding of the many issues and alternatives surrounding the "optimal" screening mammography, many of our current clinical practice guidelines are based on limited subjective assessments and anecdotal experiences, and a significant fraction is related to operational matters in busy urban environments that are staffed by experienced radiologists. The area of optimizing remote, underserved practices has been studied only in a cursory manner. Our project is but one attempt to improve our understanding of the technical, operational, and clinical issues facing these facilities and implementing technology-based solutions that may help them provide a better service to the populations they serve.

Background References:

1. S Pelikan, M Moskowitz, "Effects of lead time length bias, and false-negative assurance on screening for breast cancer," Cancer 71, 1998-2005 (1993).
2. L Tabar, G Fagerberg, HH Chen, SW Duffy, CR Smart, A Gad, RA Smith, "Efficacy of breast cancer screening by age: New results from the Swedish Two-Country Trial," Cancer 75, 2507-2517 (1995).
3. F Houn, ML Brown, "Current practice of screening mammography in the United States: Data from the national survey of mammography facilities," Radiology 190, 209-215 (1994).
4. CA Beam, PM Layde, DC Sullivan, "Variability in the interpretation of screening mammograms by US radiologists," Arch Intern Med 156, 209-213 (1996).
5. Food and Drug Administration, "Mammography facilities: requirements for accrediting bodies and quality standards and certification requirements-interim rules," Federal Register 58, 67558-72. (CFR21, Part 900) (1993).
6. JG Elmore, CK Wells, CH Lee, DH Howard, AR Feinstein, "Variability in radiologists' interpretations of mammograms," N Engl J Med 331, 1493-1499 (1994).
7. RML Warren, SW Duffy, "Comparison of single reading with double reading of mammograms and change in effectiveness with experience," Br J Radiol 68, 958-962 (1995).

8. CJ Wright, CB Mueller, "Screening mammography and public health policy: The need for perspective," Lancet **346**, 29-32 (1995).

9. LW Bassett, RE Hendrick, TL Bassford, PF Butler, D Carter, M DeBor, CJ D'Orsi, CJ Garlinghouse, RF Jones, AS Langer, JL Lichtenfeld, JR Osuch, LN Reynolds, ES de Paredes, RE Williams, "Responsibilities of the mammography facility," In: Quality determinants of mammography, clinical practice guideline. Number 13. Washington, DC: US Department of Health and Human Services, AHCPR publication no. 95-0632 (1994).

10. JG Elmore, MB Barton, VM Moceri, S Polk, PJ Arena, SW Fletcher, "Ten-year risk of false-positive screening mammograms and clinical breast examinations," N Engl J Med **338**, 1089-1096 (1998).

11. DS May, NC Lee, MR Nadel, RM Henson, DS Miller, "The National Breast and Cervical Cancer Early Detection Program: Report of the First 4 Years of Mammography Provided to Medically Underserved Women," AJR **170**, 97-104 (1998).

12. SA Feig, MJ Yaffe, "Digital mammography, computer-aided diagnosis, and telemammography," Radiol Clin North Am **33**, 1205-1228 (1995).

13. LL Fajardo, MT Yoshino, GW Seeley, R Hunt, TB Hunter, R Friedman, D Cardenas, R Boyle, "Detection of breast abnormalities on teleradiology transmitted mammograms," Invest Radiol **25**, 1111-1115 (1990).

14. MA Goldberg, "Telemammography: Implementation issues," Telemedicine Journal **1**, 215-226 (1995).

15. HK Huang, SL Lou, E Sickles, D Hoogstrate, M Jahangiri, F Cao, J Wang, "Technical issues in full-field direct digital telemammography," [Chapter] In: Computer Assisted Radiology and Surgery. Lemke HU, Inamura K, Editors. Elsevier Science B.V., 662-667 (1997).

16. HK Huang, "Digital Mammography: A Missing Link in a Totally Digital Radiology Department," Presented at the EuroPACS 97 Meeting; PISA, Italy. September 25-27, (1997).

17. JM Murphy, NJ O'Hare, D Wheat, PA McCarthy, A Dowling, R Hayes, H Bowmer, GF Wilson, MP Molloy, "Digitized mammograms: a preliminary clinical evaluation and the potential for telemammography," Journal of Telemedicine and Telecare **5**, 193-197 (1999).

18. SL Lou, HD Lin, KP Lin, D Hoogstrate, "Automatic breast region extraction from digital mammograms for PACS and telemammography applications," Computerized Medical Imaging and Graphics **24**, 205-220 (2000).

19. S Dwyer, Private communications. See also "Telemedicine Targets Mammographic Services" in Biophotonics International Nov/Dec 1997. Page 10.

20. SL Lou, EA Sickles, HK Huang, D Hoogstrate, F Cao, J Wang, M Jahangiri, "Full-field direct digital telemammography: Technical components, study protocols, and preliminary results," IEEE Trans Info Technology in Biomedicine **1**, 270-278 (1997).

21. SL Lou, HK Huang, E Sickles, D Hoogstrate, F Cao, J Wang, "Full-field direct digital telemammography: system implementation," Proc SPIE **3339**, 156-164 (1998).

22. Wu M, Zheng Y, North M, Pisano E. NLM tele-educational application for radiologists to interpret mammography. Proc AMIA Symposium, 2002, pg 909-913

23. Sheybani EO, Sankar R. ATMTN: a telemammography network architecture. IEEE Trans Biomed Eng 2002; 49:1438-1443

24. GS Maitz, TS Chang, JH Sumkin, PW Wintz, CM Johns, M Ganott, BL Holbert, CM Hakim, KM Harris, D Gur, JM Herron, "Preliminary clinical evaluation of a high-resolution telemammography System," Invest Radiol **32**, 236-240 (1997).
25. JM Holbert, M Staiger, TS Chang, JD Towers, CA Britton, "Selection of processing algorithms for digital image compression: A rank-order study," Acad Radiol **2**, 273-276 (1995).

Appendix

See Attached.

1 - 7

Subjective assessment of high-level image compression of digitized mammograms

J. Ken Leader*[a], Jules H. Sumkin[ab], Marie A. Ganott[ab], Christiane Hakim[ab], Lara Hardesty[ab], Ratan Shah[ab], Luisa Wallace[ab], Amy Klym[a], John M. Drescher[a], Glenn S. Maitz[a], David Gur[a]

[a]University of Pittsburgh, Pittsburgh, PA USA 15213
[b]Magee-Womens Hospital, Pittsburgh, PA USA 15213

ABSTRACT

This study was designed to evaluate radiologists' ability to identify highly-compressed, digitized mammographic images displayed on high-resolution, monitors. Mammography films were digitized at 50 micron pixel dimensions using a high-resolution laser film digitizer. Image data were compressed using the irreversible (lossy), wavelet-based JPEG 2000 method. Twenty images were randomly presented in pairs (one image per monitor) in three modes: mode 1, non-compressed versus 50:1 compression; mode 2, non-compressed versus 75:1 compression; and mode 3, 50:1 versus 75:1 compression with 20 random pairs presented twice (80 pairs total). Six radiologists were forced to choose which image had the lower level of data compression in a two-alternative forced choice paradigm. The average percent correct across the six radiologists for modes 1, 2 and 3 were 52.5% (+/-11.3), 58.3% (+/-14.7), and 58.3% (+/-7.5), respectively. Intra-reader agreement ranged from 10 to 50% and Kappa from -0.78 to -0.19. Kappa for inter-reader agreement ranged from -0.47 to 0.37. The "monitor effect" (left/right) was of the same order of magnitude as the radiologists' ability to identify the lower level of image compression. In this controlled evaluation, radiologists did not accurately discriminate non-compressed and highly-compressed images. Therefore, 75:1 image compression should be acceptable for review of digitized mammograms in a telemammography system.

Keywords: Image compression, data compression, JPEG 2000, telemammography

1. INTRODUCTION

Breast cancer screening mammography is widely practiced and increasingly challenging to manage in the clinical environment, but there is potential for improvement.[1-7] Teleradiology is an approach that may provide more timely patient management. Image compression,[8-13] image cropping,[12-14] and image selection[15] are commonly used in teleradiology to facilitate the timely transmission of data. The high-spatial resolution required for mammography complicates the design and implementation of a telemammography system. The large mammographic image file size (33-55 MBytes per image) is one obstacle to timely transmission of data, especially across low-level data connections. High-level image compression may assist in overcoming this obstacle and can only be realized with lossy image compression techniques, which necessitates the loss of some image information and a degree of image degradation.

The use of high-level image compression in medical applications is frequently met with skepticism because of the potential degradation of the depiction of objects under investigation. Human observer performance studies designed to evaluate wavelet compression of medical images for clinical applications have reported acceptable compression levels ranging from 8:1 to 100:1.[16-26] Wavelet-based compression, the trend in medical image compression, is reported to be superior to the original JPEG compression based on the direct cosine transform in terms of image quality at high-levels of image compression.[16,17]

---

* jklst3@pitt.edu; phone (412) 641-2572; fax (412) 641-2582, University of Pittsburgh, Magee-Womens Hospital, 300 Halket Street, Suite 4200, Pittsburgh, PA 15213

From our perspective the effect of image degradation from lossy compression of medical image interpretation remains unresolved, particularly regarding mammography. Observer studies reported that 8:1[22] and 10:1[27] compression ratios are acceptable for mammography applications using both wavelet and the original JPEG compression methods. Visualization of calcifications depicted on digitized mammograms was subjectively rated as excellent for wavelet compression ratios as high as 56:1.[19] Uncompressed digitized mammographic images were rated to be comparable to images compressed at 30:1 using wavelet compression.[20] These studies are indeed promising, and high-levels of image compression may be ultimately clinically acceptable in mammography.

Powell et al.[22] (2000) conducted a clinical evaluation that compared film mammography to digitized images compressed at 8:1 using wavelet based compression. The accuracy for detecting malignancy was not statistically different when depicted on film or digitized images in a receiver operating characteristics (ROC) study. The false positive rate at a fixed sensitivity of 0.90 was significantly lower (better) using digitized images as compared with film. Compressed digitized images were also slightly better (though not statistically) than film in terms of recall rate for negative mammograms and those depicting benign findings. The recall rate for mammograms depicting malignant abnormalities was slightly better (though not statistically) when original films were used as compared with digitized images.

The objective of this study was to determine an acceptable level of image compression in a telemammography application. The ability of radiologists to discriminate high-levels of image compression as applied to digitized mammograms was evaluated. Image pairs of different compression levels were randomly presented and viewed side-by-side on two high-resolution monitors. Six radiologists were forced to choose the lower level of image compression and rate the relative utility of the images for use in a screening mammography environment.

2. METHODS

2.1 Case selection

This study used twenty breast cancer screening examinations randomly selected from a larger telemammography project, which was designed to evaluate the ability telemammography to reduce the number of patients being recalled for additional imaging procedures. One image view from each case (i.e., twenty images total) was selected to represent each examination. The verified findings depicted in these examinations included masses and calcification clusters (Table 1). The dataset for this retrospective study was assembled and analyzed under University of Pittsburgh Institutional Review Board approved protocol, and the image data was anonymized.

Table 1
Image views and depicted abnormalities

View	Mass	Calcifications	Mass & calcifications	No finding
MLO	3	2	2	3
CC	3	3	1	3

MLO - mediolateral oblique
CC - craniocaudal

2.2 Image processing

Mammographic films were digitized at 50 micron pixel dimensions and 12-bit grayscale using a high-resolution, laser film digitizer (Lumiscan 85, Eastman Kodak, Rochester, NY, USA). Each digitized mammographic image was automatically cropped to decrease the non-tissue area surrounding the breast. The cropped image data were compressed using the irreversible (lossy), 9/7 transform, wavelet-based JPEG 2000 method at compression ratios of 50:1 and 75:1 and subsequently decompressed prior to display. A total of sixty images were generated for the study, the twenty original digitized images plus two compressed images at 50:1 and 75:1 ratios for each of these (or a total of sixty images).

2.3 Image display

The images were displayed on two calibrated, high-resolution (2048 x 2560), 8-bit grayscale, portrait monitors at a nominal setting of 80 ftL (DS5100P, Clinton Electronics, Rockford, IL, USA). Typically, when a single image displayed on the monitor the display scale was approximately 100 micron per pixel. Minimal unsharp masking was employed. In short, image data were first smoothed with a 2-D 129 mean kernel, and subsequently the weighted (0.10) smoothed image was subtracted from the original image. Finally, the resulting pixel values were re-scaled from 0 to 4095. Image magnification and window/level adjustments were not permitted during the study.

Fixed look-up table (LUT) values are automatically calculated based on the pixel value distribution (histogram). In short, the typical pixel value distribution of digitized mammographic images is bimodal. The center between the two modes was set as the level value (brightness), and the span of the two modes was set as the window value (contrast). Additionally, the cropped images were padded (filled) prior to display to restore the full height of the image.

2.4 Study protocol

Six experienced radiologists participated in the study. They were presented image pairs (one image per monitor) that consisted of the same image at different levels of compression (Fig. 1). The images were paired in three modes: mode 1, non-compressed versus 50:1 compression; mode 2, non-compressed versus 75:1 compression; and mode 3, 50:1 versus 75:1 compression. The sixty image pairs were randomly presented with 20 randomly selected pairs presented a second time to evaluate intra-observer variability (or a total of eighty pairs). Compression levels were also randomly assigned between the two monitors for counterbalancing.
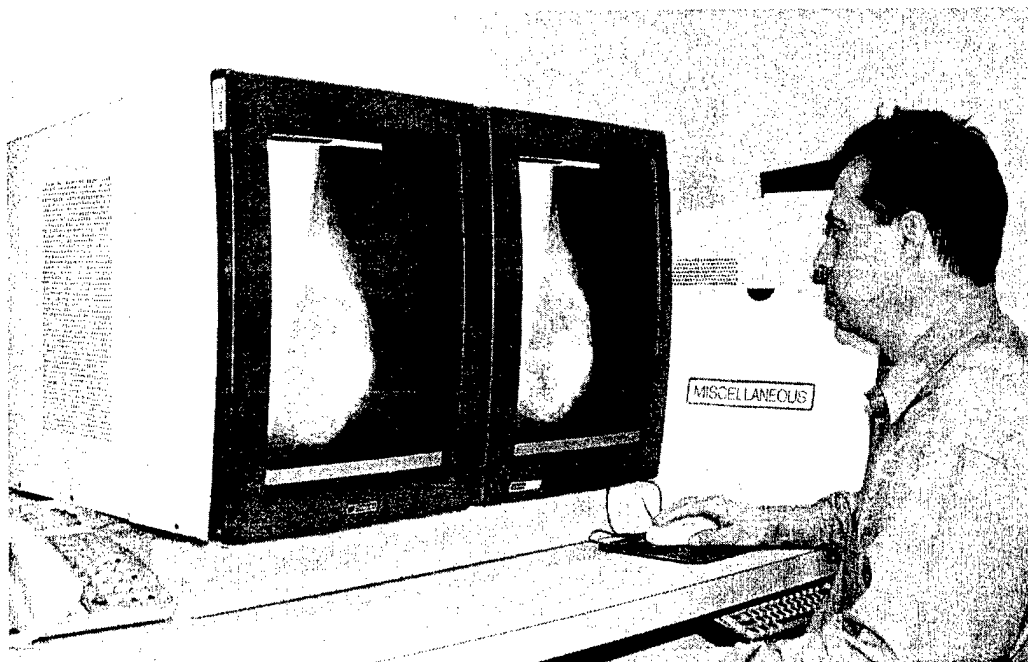

Fig. 1. Telemammography workstation used for the study.

In a 2-AFC paradigm the radiologists were forced to choose the image (i.e., right or left monitor) that had the lower level of data compression. In addition, they compared and rated the clinical utility between the two images presented in each pair. After image review, two questions were presented on a computer scoring form and answered using the computer mouse (Fig. 2). The radiologists were given written instructions regarding the protocol:

You will be presented with 80 pairs of images, one image on each monitor. The window and level values for the monitor display will be fixed. Magnification features will not be available during this study. One image will contain less information than the other as a result of data compression. The monitor that displays the less compressed image will be randomly selected. The same image pairs will appear multiple times throughout the study. After you have reviewed the images, the "eval case" button on the bottom task bar will bring up two questions to be answered.

Fig. 2 Computerized scoring form completed for each image pair.

2.5 Data analysis

The average percent correct decisions across the six readers for discriminating the lower level of image compression was compared with a random (chance) selection using a one-sample T-test for each mode and each monitor. Friedman Two-Way Analysis of Variances by Ranks was used to test if there was a difference between modes. Kappa was used to evaluate intra-reader agreement for the twenty repeated pairs of images and inter-reader agreement for each mode. To determine if a learning effect was present the percent correct decision for the first, second, and third presentations of pairs of images was tested for trend using the Page Test for Ordered Alternatives. All images were presented a minimum of three times with the twenty repeated pairs randomly selected. The percent of image pairs rated as clinically equivalent for both the correct or incorrect decisions for identifying the lower level of image compression were compared to random (chance) selection using a one-sample T-test for each mode and each monitor.

3. RESULTS

The subjective appearance of the compressed images was extremely similar to the original uncompressed image. The task of discriminating the more compressed image in each pair was reported to be difficult by all readers. The smoothing effect of wavelet compression did not produce distinguishable image features such as blocking artifacts characteristic of high-level original JPEG compression.

Readers' ability to correctly discriminate the lower level of image compression was only slightly better than chance and was of the same order of magnitude as the "monitor effect" (Table 2). Readers' performance levels were not significantly different across the three presentation modes ($p > 0.05$). However, the readers correctly identified images compressed at 50:1 ratio as lower than 75:1 image compression at a rate significantly greater than chance ($p < 0.05$). On average the readers performed better when the lower level of compression was presented on the left monitor for all three modes, but the "monitor effect" (left versus right) was not significant.

Table 2
Average percent correct for discriminating the lower compression level for all image pairs when the correct image was on the right monitor and the left monitor

	mode 1[ad]	mode 2[b]	mode 3[c]
All images	52.5 (11.3)	58.3 (14.7)	58.3 (7.5)[e]
Images on right monitor	45.7 (25.3)	43.2 (25.8)	47.8 (26.5)
Images on left monitor	62.5 (14.1)	73.2 (25.3)	69.0 (23.1)

[a] mode 1 - non-compressed & 50:1 compression
[b] mode 2 - non-compressed & 75:1 compression
[c] mode 3 - 50:1 & 75:1 compression
[d] group mean and standard deviation in ( )
[e] $p < 0.05$ one sample T-test

Intra- and inter-reader agreements for discriminating the lower level of data compression were poor for the individual as well as between readers (Tables 3 and 4). Kappa for intra-reader agreement for readers 1, 2, 3, 4, 5, and 6 were -0.25, -0.39, -0.30, -0.19, -0.78, and -0.30, respectively. No two readers consistently agreed across the three presentation modes. Inter-reader Kappa for discriminating the lower level of image compression for the six readers ranged from -0.47 to 0.26, -0.36 to 0.37, and -0.30 to 0.30 for modes 1, 2, and 3, respectively (Table 4).

Table 3
Comparison between the first and second reads of the twenty repeated image pairs

reader	first read	second read correct[a]	second read incorrect
1	correct	10 (2)	30 (6)
	incorrect	30 (6)	30 (6)
2	correct	15 (3)	30 (6)
	incorrect	40 (8)	15 (3)
3	correct	20 (4)	30 (6)
	incorrect	35 (7)	15 (3)
4	correct	5 (1)	25 (5)
	incorrect	25 (5)	45 (9)
4	correct	5 (1)	40 (8)
	incorrect	50 (10)	5 (1)
5	correct	10 (2)	50 (10)
	incorrect	20 (4)	20 (4)

[a] percentage and number in ( )

Table 4
Kappa for inter-reader agreement for the six readers and the three presentation modes

mode	reader	reader 2	3	4	5	6
1[a]	1	-0.471	-0.042	0.043	-0.200	-0.038
	2		0.118	0.223	-0.100	-0.237
	3			0.255	-0.200	0.151
	4				-0.100	-0.101
	5					-0.300
2[b]	1	0.175	-0.359	-0.354	-0.300	0.368
	2		-0.284	-0.023	0.100	-0.177
	3			0.018	0.100	-0.217
	4				-0.200	0.125
	5					-0.300
3[c]	1	-0.099	-0.300	0.121	0.100	-0.237
	2		-0.100	-0.099	0.100	0.175
	3			0.100	-0.200	0.100
	4				0.300	-0.031
	5					-0.100

[a] mode 1 - non-compressed & 50:1 compression
[b] mode 2 - non-compressed & 75:1 compression
[c] mode 3 - 50:1 & 75:1 compression

A slight learning effect was observed in the average reader's ability to select the lower level of image compression during the first three presentations (Table 5). The mean percent for correctly discriminating the lower level of image compression showed an increasing trend across the three presentations that was not significant ($p > 0.05$). Reader 6 was an outlier, and, although the trend was not significant, excluding this reader from the analysis removed the increasing trend across the three presentations.

Table 5
Percent correct for selecting the less compressed image during the first, second, and third presentations

reader	first (n= 20)	second (n = 20)	third (n = 20)
1	65.0	50.0	60.0
2	55.0	55.0	60.0
3	50.0	50.0	55.0
4	60.0	65.0	70.0
5	50.0	50.0	50.0
6	35.0	80.0	65.0
mean	52.5	58.3	60.0[a]
std	10.4	12.1	7.1

[a] $p > 0.05$

Images correctly identified as less compressed by the readers were rated as "clinically equivalent" at relatively the same rate as images incorrectly identified (Table 6). However, on the left monitor the readers rated correctly selected images as "clinically equivalent" more often than random selection ($p < 0.05$). The average number of image pairs rated as clinically equivalent by the six radiologist were 14.2 ($\pm$ 4.8), 14.2 ($\pm$ 4.1), and 13.3 ($\pm$ 5.5) out of the twenty possible pairs for modes 1, 2, and 3, respectively.

Table 6
Percent of image pairs rated "clinically equivalent" for correct and incorrect selection of lower compression level for either monitor, the right monitor, and the left monitor

	correct choice of lower compression level			incorrect choice of lower compression level		
mode	either monitor[d]	right monitor	left monitor	either monitor	right monitor	left monitor
1[a]	48.3 (20.1)	24.0 (13.6)	24.3 (15.2)	51.7 (20.1)	34.2 (24.9)	17.5 (10.5)
2[b]	62.3 (19.2)	18.9 (15.6)	43.4 (17.8)[e]	37.7 (19.2)	26.3 (15.9)	11.4 (12.9)[e]
3[c]	53.1 (18.6)	19.2 (15.4)	33.9 (17.0)	46.9 (18.6)	27.1 (20.7)	19.8 (14.7)

[a] mode 1 - non-compressed & 50:1 compression
[b] mode 2 - non-compressed & 75:1 compression
[c] mode 3 - 50:1 & 75:1 compression
[d] group mean and standard deviation in ( )
[e] $p < 0.05$ one sample T-test

4. DISCUSSION

In this controlled evaluation, image compression achieved with wavelet-based JPEG 2000 was not reliably discriminated and rated by radiologists and, therefore, could be considered applicable for telemammography applications. Radiologists did not accurately or reliably select the lower level of image compression between image pairs when presented side-by-side with non-compressed images and those compressed at 50:1 and 75:1 compression levels. Interestingly, the "monitor effect" (left versus right) was of the same order of magnitude as the radiologists' ability to discriminate the lower level of image compression. As a group the readers' ability to identify the lower level of data compression slightly improved across the readings, but not significantly. The majority of image pairs, which were compressed at different ratios, were rated as "clinically equivalent" for use in a screening environment independent of whether the readers selected correctly or incorrectly the less compressed image.

The images in our study were presented on separate, side-by-side monitors with magnification, pan zoom, and window/level features disabled. Permitting magnification and window/level may (or may not) have improved discrimination. A similar 2-AFC study by Slone et al.[17] (2000) evaluated wavelet and original JPEG compression of

posteroanterior chest digital radiographs and reported that image degradation was detected at compression levels greater than 11:1 for both compression methods. At a compression level of 75:1 the lower compressed image was correctly identified approximately 95 % of the time for both the wavelet and the JPEG compression methods. The images were presented on a single monitor, and the readers were permitted to magnify and toggle between images, which they acknowledged was conservative and tested the reader's temporal sensitivity.

Since radiologists could not accurately or reliably discriminate non-compressed and highly-compressed mammographic images, their interpretation using either non-compressed or highly-compressed images is not likely to differ substantially. We also note that diligent monitor calibration may be critical to image fidelity.

ACKNOWLEDGEMENTS

REFERENCES

1. Chamot E and Perneger TV. "Misconception about efficacy of mammography screening: a public health dilemma." *J Epidemiol Community Health* **55**(11):799-803, 2001.
2. Coughlin SS, Thompson TD, Hall HI, Logan P, Uhler RJ. Breast and cervical carcinoma screening practices among women in rural and nonrural areas of the United States, 1998-1999. *Cancer* **94**(11):2801-2812, 2002.
3. Michaelson J, Satija S, Moore R, Weber G. Halpern E, Garland A, Puri D, Kopans DB. "The pattern of breast cancer screening utilization and its consequences." *Cancer* **94**(1):37-43, 2002.
4. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. "Variability in radiologists' interpretations of mammograms." *N Engl J Med* **331**(22):1493-1499, 1994.
5. Warren RML and Duffy SW. "Comparison of single reading with double reading of mammograms and change in effectiveness with experience." *Br J Radiol* **68**(813):958-962, 1995.
6. Hulka CA, Slanetz PJ, Halpern EF, Hall DA, McCarthy KA, Moore R, Boutin S, Kopans DB. "Patients' opinion of mammography screening services: immediate results versus delayed results due to interpretation by two observers." *AJR Am J Roentgenol* **168**:1085-1089, 1997.
7. Yawn B, Krein S, Christianson J, Hartley D, Moscovice I. "Rural radiology: who is producing images and who is reading them?" *J Rural Health* **13**(2):136-144, 1997.
8. Bolle SR, Sund T, Stormer J. "Receiver operating characteristic study of image processing for teleradiology and digital workstations." *J Digit Imaging* **10**(4):152-157, 1997.
9. Maitz GS, Chang TS, Sumkin JH, Wintz PW, Johns CM, Ganott M, Holbert BL, Hakim CM, Harris KM, Gur D, Herron JM. "Preliminary clinical evaluation of a high-resolution telemammography System." *Invest Radiol* **32**(4):236-240, 1997.
10. Mitra S, Yang S, Kustov V. "Wavelet-based vector quantization for high-fidelity compression and fast transmission of medical images." *J Digit Imaging* **11**(4 Suppl 2):24-30, 1998.
11. Kalyanpur A, Neklesa VP, Taylor CR, Daftary AR, Brink JA. "Evaluation of JPEG and wavelet compression of body CT images for direct digital teleradiology transmission." *Radiology* **217**(3):772-779, 2000.
12. Drescher JM, Maitz GS, Leader JK, Sumkin JH, Poller WR, Klaman H, Zheng B, Gur D. "Design considerations for a multi-site, POTS-based telemammography system." *Proceedings of SPIE Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation* **4685**:416-421, February 2002.
13. Drescher JM, Maitz GS, Traylor C, Leader JK, Clearfield RJ, Shah R, Ganott MA, Pugliese F, Duffner D, Lockhart J, Gur D. "A multi-site telemammography system: preliminary assessment of technical and operational issues." *Proceedings of SPIE Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, **5033**:360-369, February 2003.
14. Lou SL, Lin HD, Lin KP, Hoogstrate D. "Automatic breast region extraction from digital mammograms for PACS and telemammography applications." *Comput Med Imaging Graph* **24**(4):205-220, 2000.
15. Ludwig K, Bick U, Oelerich M, Schuierer G, Puskas Z, Nicolas K, Koch A, Lenzen H. "Is image selection a useful strategy to decrease the transmission time in teleradiology? A study of 100 emergency cranial CTs." *Eur Radiol* **8**(9):1719-1721, 1998.

16. Ricke J, Maass P, Hänninen EL; Liebig T, Amthauer H, Stroszczynski C, Schauer W, Boskamp T, Wolf M. "Wavelet versus JPEG (Joint Photographic Expert Group) and fractal compression: impact on the detection of low-contrast details in computed radiographs." *Invest Radiol* 33(8):456-463, 1998.

17. Slone RM, Foos DH, Whiting BR, Muka E, Rubin DA, Pilgram TK, Kohm KS, Young SS, Ho P, Hendrickson DD. "Assessment of visually lossless irreversible image compression: comparison of three methods by using an image-comparison workstation." *Radiology* 215:543-553, 2000.

18. Goldberg MA, Pivovarov M, Mayo-Smith WW, Bhalla MP, Blickman JG, Bramson RT, Boland GW, Llewellyn HJ, Halpern E. "Application of wavelet compression to digitized radiographs." *AJR Am J Roentgenol* 163:463-468, 1994.

19. Lucier BJ, Kallergi M, Qian W, DeVore RA, Clark RA, Saff EB, Clarke LP. "Wavelet compression and segmentation of digital mammograms." *J Digit Imaging* 7(1):27-38, 1994.

20. Perlmutter SM, Cosman PC, Gray RM, Olshen RA, Ikeda D, Adams CN, Betts BJ, Williams MB, Perlmutter KO, Li J, Aiyer A, Fajardo L, Birdwell R, Daniel BL. "Image quality in lossy compressed digital mammograms." *Signal Processing* 59:189-210, 1997.

21. Persons KR, Palisson PM, Manduca A, Charboneau WJ, James EM, Charboneau NT, Hangiandreou NJ, Erickson BJ. "Ultrasound grayscale image compression with JPEG and wavelet techniques." *J Digit Imaging* 13(1):25-32, 2000.

22. Powell KA, Mallasch PG, Obuchowski NA, Kerczewski RJ, Ganobcik SN, Cardenosa G, Chilcote W. "Clinical evaluation of wavelet-compressed digitized screen-film mammography." *Acad Radiol* 7(5):311-316, 2000.

23. Terae S, Miyasaka K, Kudoh K, Nambu T, Shimizu T, Kaneko K, Yoshikawa H, Kishimoto R, Omatsu T, Fujita N. "Wavelet compression on detection of brain lesions with magnetic resonance imaging." *J Digit Imaging* 13(4):178-179, 2000.

24. Trapnell CJ, Scarfe WC, Cook JH, Silveira AM, Regennitter FJ, Haskell BS. "Diagnostic accuracy of film-based, TIFF, and wavelet compressed digital temporomandibular joint images." *J Digit Imaging* 13(1):38-45, 2000.

25. Zheng LM, Sone S, Itani Y, Wang Q, Hanamura K, Asakura K, Li F, Yang ZG, Wang JC, Funasaka T. "Effect of CT digital image compression on detection of coronary artery calcification." *Acta Radiol* 41(2):116-121, 2000.

26. Zalis ME, Hahn PF, Arellano RS, Gazelle GS, Mueller PR. "CT colonography with teleradiology: effect of lossy wavelet compression on polyp detection-initial observations." *Radiology* 220:387-392, 2001.

27. Good WF, Maitz GS, Gur D. "Joint Photographic Experts Group (JPEG) compatible data compression of mammograms." *J Digit Imaging* 7(3):123-132, 1994.

Computer-Aided Detection Schemes: The Effect of Limiting the Number of Cued Regions in Each Case

Bin Zheng[1]
Joseph K. Leader
Gordon Abrams
Betty Shindel
Victor Catullo
Walter F. Good
David Gur

OBJECTIVE. We assessed performance changes of a mammographic computer-aided detection scheme when we restricted the maximum number of regions that could be identified (cued) as showing positive findings in each case.

MATERIALS AND METHODS. A computer-aided detection scheme was applied to 500 cases (or 2,000 images), including 300 cases in which mammograms showed verified malignant masses. We evaluated the overall case-based performance of the scheme using a free-response receiver operating characteristic approach, and we measured detection sensitivity at a fixed false-positive detection rate of 0.4 per image after gradually reducing the maximum number of cued regions allowed for each case from seven to one.

RESULTS. The original computer-aided detection scheme achieved a maximum case-based sensitivity of 97% at 3.3 false-positive detected regions per image. For a detection decision score set at 0.565, the scheme had a 79% (237/300) case-based sensitivity, with 0.4 false-positive detected regions per image. After limiting the number of maximum allowed cued regions per case, the false-positive rates decreased faster than the true-positive rates. At a maximum of two cued regions per case, the false-positive rate decreased from 0.4 to 0.21 per image, whereas detection sensitivity decreased from 237 to 220 masses. To maintain sensitivity at 79%, we reduced the detection decision score to as low as 0.36, which resulted in a reduction of false-positive detected regions from 0.4 to 0.3 per image and a reduction in region-based sensitivity from 66.1% to 61.4%.

CONCLUSION. Limiting the maximum number of cued regions per case can improve the overall case-based performance of computer-aided detection schemes in mammography.

omputer-aided detection systems are routinely used in a number of medical institutions around the world to assist radiologists in the detection of abnormalities depicted on mammograms. The number of mammograms scanned through commercial computer-detection systems has been rapidly increasing. Although no general agreement has been reached on how computer-aided detection affects radiologists' performance in terms of sensitivity and specificity [1–4], there are indications that the performance of the computer-aided detection scheme itself has an impact on radiologists' performance in detecting abnormalities [5, 6], and observer confidence levels in accepting the cues generated by these systems increases with higher performance levels of the scheme [7, 8]. Several commercial computer-aided detection systems have been approved by the United States Food and Drug Administration, and the relative per-formance levels of such systems have been compared [9, 10]. All commercial computer-aided detection systems use specific threshold values to determine whether an identified suspicious region is ultimately cued as a positive finding, and the performance of these systems is frequently evaluated on the basis of the case-based sensitivity achieved at a given false-positive detection rate. In a case-based (or a breast-based) analysis, sensitivity is based on the correct detection of at least one true-positive region on either the craniocaudal or mediolateral oblique mammographic view or on both [1].

Evaluation of computer-aided detection performance is not a simple matter. Previous studies have shown that performance can vary widely depending on which scoring method is used, and there is no general agreement on which scoring method should be used for this purpose [11, 12]. One study showed that at approximately the same false-positive rate (e.g., 1.5 per image), the

measured sensitivity for the detection of micro-calcification clusters ranged between 45% and 85% depending on which of three different assessment methods were used [11].

In addition, computer-aided detection performance depends on the composition of the image database used [13]. In general, computer-aided detection schemes may identify a large number of suspicious regions on some images (e.g., images depicting dense tissue patterns), but only a few suspicious regions on other images (e.g., images dominated by fatty tissue) [14]. Therefore, limiting the maximum number of suspicious regions allowed to be cued for one case could potentially reduce the false-positive rate with a relatively small decrease in sensitivity. This approach is used in commercially available systems, but to the best of our knowledge, the effect of implementing the approach on image- and case-based sensitivity and false-positive detection rates has not been described in detail. This study was performed to assess this issue.

Materials and Methods

We selected 500 cases (or 2,000 digitized mammograms) from a large image database available in our laboratory. Among these cases, verified malignant masses were depicted in 300 cases, and the remaining 200 were negative findings. In all cases with positive findings, a panel of radiologists identified the locations of the mass regions on the images using the original diagnostic and biopsy reports. The central coordinates (x and y) of each mass region

were visually identified, marked, and saved in a "truth file." In this data set, mass regions were visible on both the craniocaudal and mediolateral oblique mammographic views in 270 cases and were only visible on one of the two views in 30 cases. Thus, 570 mass regions were identified on the images in this study. Figure 1 shows the size distribution of the 300 masses in the data set.

A computer program determined the size of each mass region by counting the total number of pixels inside the identified boundary contour of the region (multiplied by 0.0016 cm^2 per pixel). The size of a mass was represented by a large computed area on either the craniocaudal or mediolateral oblique mammogram. For each identified mass region, the panel of radiologists assigned a subjective rating of subtlety using a 5-point rating scale that ranged from 1 (very easily visible) to 5 (very subtly visible). Figure 2 shows the distribution of assigned subtlety ratings in this data set. Subtlety of a mass was represented by the lower rating assigned to either the craniocaudal or mediolateral oblique mammographic view. We verified all cases with negative (or benign) findings by reviewing the available diagnostic information and the data from a follow-up examination with negative results, confirming a minimum of one disease-free year.

A computer-aided detection scheme developed previously in our laboratory [15] was applied to the 2,000 images in the data set. Because we only examined computer-aided detection performance for mass detection in this study, each image was first reduced by pixel averaging (a factor of 8 in both x and y directions), increasing the effective pixel size from 50 × 50 μm in the original digitized image to 400 × 400 μm. The mass detection scheme then identified between 10 and 30 suspicious regions in each image depending on the regional tissue patterns. For each identified region,

a multilayer regional growth algorithm [16] was applied to define the contours of the region as depicted in the image. If the region met simple growth criteria, a set of features from the interior and surrounding background of the region was computed by the scheme. Otherwise, the region was considered to have negative findings and was deleted. Finally, a feature-based artificial neural network classified each suspicious region as showing positive or negative findings by assigning a detection (or probability) score. In a manner similar to the commercial computer-aided detection products, our detection scheme identified a region as having a positive finding if the detection score exceeded a predetermined threshold. If the detection score did not exceed the threshold, the region was not cued and was considered to be a negative finding.

After processing all images, we compared the regions with detected positive findings with the results saved in the truth file. To determine whether a detected region was considered a true-positive finding, we applied the following criterion: If the distance between the computed center of a detected region and the visually marked coordinate on a mammogram was shorter than the effective radius (the average radial length computed by the computer-aided detection scheme), the region was considered to be a match to a true-positive mass. Otherwise, the region was considered a false-positive case.

To show the original performance of the computer-aided detection scheme when applied to this data set, we plotted free-response receiver operating characteristic curves for both case-based and region-based scores. In the case-based performance curve, sensitivity was assessed on the basis of the correct marking of at least one true-positive region in either (or both) of the two mammographic views, and if two regions were detected, the higher score was se-
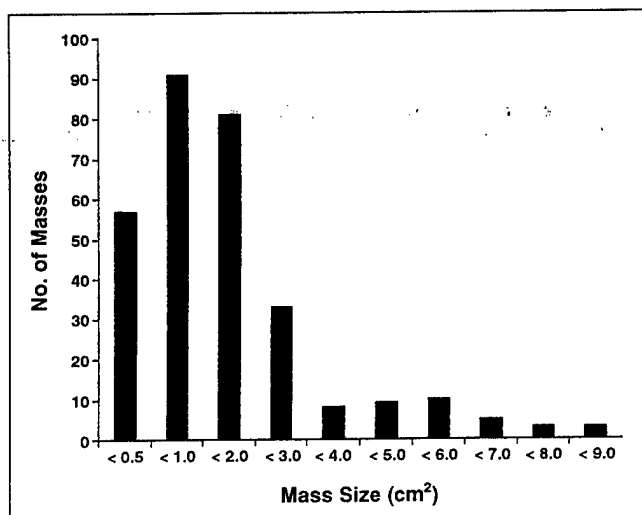

Fig. 1.—Bar graph shows size distribution of 300 masses depicted in data set. Mass size is represented by larger depicted area on either craniocaudal or mediolateral oblique mammographic view.
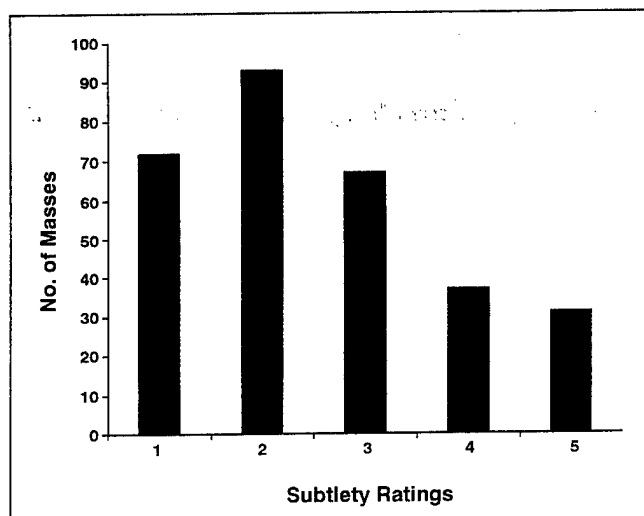
Fig. 2.—Bar graph shows distribution of subjectively rated subtlety of 300 masses depicted in data set. Subtlety of each identified mass was rated on 5-point scale, ranging from 1 (very easily visible) to 5 (very subtly visible). Mass subtlety is represented by lower-rated depiction on either craniocaudal or mediolateral oblique mammographic view.

lected to represent the mass. In the region-based performance curve, if the same mass was depicted on both craniocaudal and mediolateral oblique views, we considered these two images to represent two independent regions.

We applied a threshold score to the artificial neural network results to evaluate the sensitivity of the scheme at different false-positive rates. We also adjusted the threshold value to produce a false-positive rate comparable to that of the leading commercial computer-aided detection systems (e.g., a false-positive rate of 0.4 regions per image [2]). By changing the total number of cued regions permitted in each case to anywhere from seven to one, we compared the change in performance levels (including both sensitivity and false-positive rate). The scores generated by the artificial neural networks for all detected regions were sorted by value from the highest to the lowest, and the regions with higher scores were selected sequentially until the predetermined limit of cued regions per case was reached. In addition, we kept the case-based sensitivity constant by reducing the detection threshold and assessed the changes in false-positive rates and image-based sensitivity as the total number of allowed cues per case was reduced from seven to two.

Results

Figure 3 shows two computed free-response receiver operating characteristic curves after the application of our computer-aided detection scheme to this data set. One is a case-based free-response receiver operating characteristic performance curve; the other is a region-based curve. Setting the threshold value of the artificial neural network detection scores at 0.565 generated a decision threshold line, as shown in Figure 3. At this level, the computer-aided detection scheme identified 79% of the malignant masses with 0.4 false-positive regions per image being cued. At this threshold, the scheme did not detect any false-positive regions in 33.2% (166/500) of the cases.

Table 1 provides the performance levels of the computer-aided detection scheme when we limited the maximum number of cued regions allowed in one case at this threshold level (0.565). The false-positive detection rate decreased substantially faster than the case-based sensitivity. For example, when we limited the maximum number of cued regions to two per case, the detection sensitivity decreased by 7.2% (from 237/300 to 220/300 cases), whereas the false-positive detection rate decreased by 47.3% (from 0.40 to 0.21 per image). In 65% of the true-positive cases, the region with the highest artificial neural network score was the malignant mass region (Table 1).

Figure 4 shows five free-response receiver operating characteristic curves generated when
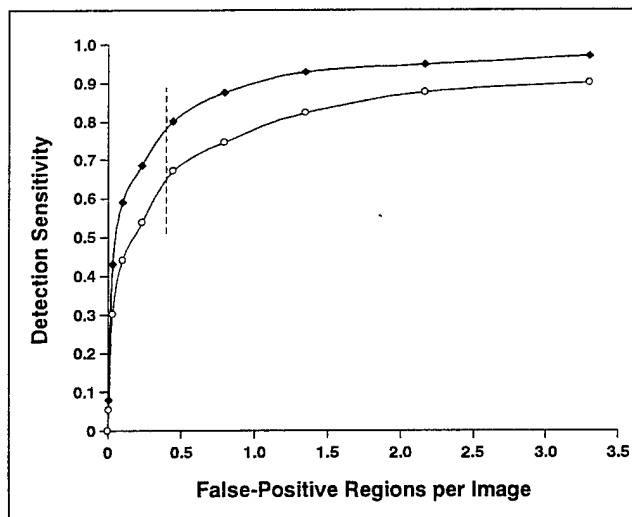
the maximum allowed number of cues per case was limited to between seven and two. As the maximum number of allowed cues was reduced, the free-response receiver operating characteristic curves tended to become steeper. Table 2 summarizes the results after limiting the maximum number of cued regions and changing the threshold value of the artificial neural network detection scores to maintain a 79% case-based sensitivity. The table shows that we were able to reduce the false-positive rates while maintaining a constant sensitivity.

For example, by limiting the maximum allowed number of cues to two per case and adjusting the artificial neural network threshold to 0.36, we reduced the false-positive rate from 0.4 to 0.3 regions per image.

One interesting finding was that the 17 (of the 237) masses detected using these two scoring methods were not identical. When the maximum number of cued regions was limited to two per case, 17 masses with artificial neural network scores higher than 0.565 (range, 0.57–0.77) were eliminated. Reducing the

Fig. 3.—Graph illustrates overall performance of computer-aided detection scheme when applied to database of 2,000 mammograms (500 cases) with no limitation on number of cued regions. Detection decision threshold line is represented by dotted line. ◆ = case-based free-response receiver operating characteristic curve, O = image-based free-response receiver operating characteristic curve.

TABLE 1	Performance Levels of Computer-Aided Detection as a Function of the Maximum Number of Cued Regions Allowed per Case					
Maximum No. of Cued Regions Allowed per Case	Sensitivity[a]				False-Positive Regions[b]	
	Case-Based		Region-Based			
	No.[c]	%	No.[d]	%	No.	Per-Image Rate
No limit	237	79.0	377	66.1	803	0.40
7	237	79.0	376	66.0	795	0.40
5	236	78.7	370	64.9	753	0.38
4	233	77.7	364	63.9	695	0.35
3	227	75.7	351	61.6	588	0.29
2	220	73.3	316	55.4	423	0.21
1	195	65.0	195	34.2	224	0.11

Note.—Artificial neural network threshold vaue was set at 0.565.
[a]Detected true-positive cases.
[b]Detected false-positive regions.
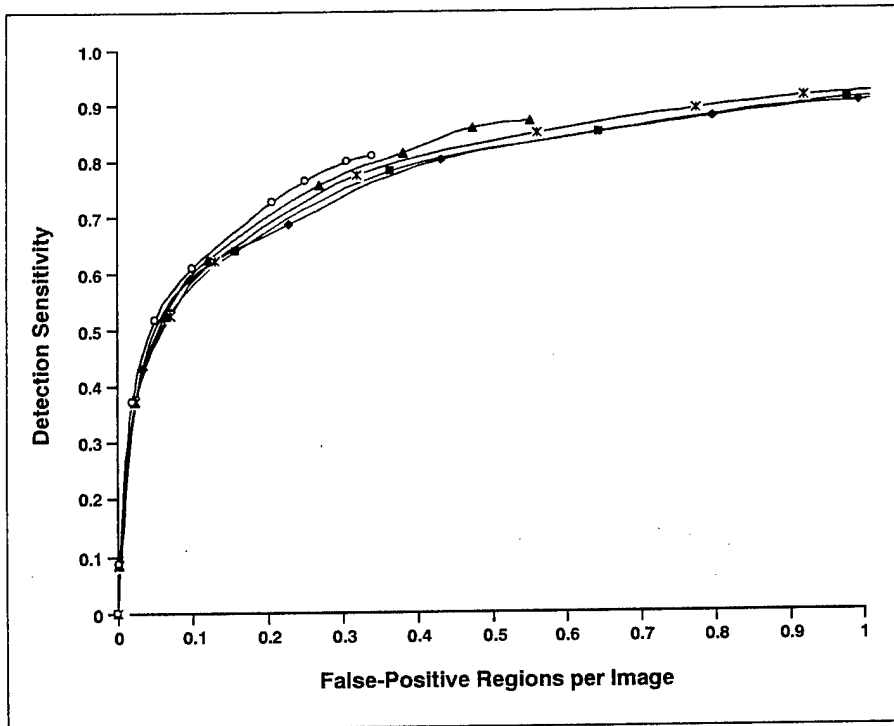[c]Cases.
[d]Regions.

Fig. 4.—Graph shows five plots depicting free-response receiver operating characteristic curves generated by different maximum numbers of cued regions allowed per case. Maximum number of cued regions indicated by ◆ = no limit, ■ = ≤ 7, ✕ = ≤ 5, ▲ = ≤ 3, ○ = ≤ 2.

threshold score to 0.36 resulted in the identification of 17 different masses with artificial neural network scores in the range between 0.36 and 0.51. Figure 5 shows the distribution of mass sizes and subtlety ratings of the 34 masses missed by both scoring methods. The results suggest that the 17 masses that were detected only when the number of allowed cues was limited to two per case and the threshold was lowered tended to be somewhat small. All 34 masses were actually positive findings. At this time, the follow-up period on these patients has not been long enough to assess the difference (if any) in clinical impact of the two approaches.

Discussion

Case distributions and rating methods could have a significant effect on the evaluation of computer-aided detection performance levels [11–13]. In this study, we tested a simple scoring method that alters measured performance. The method of limiting the maximum number of cued regions allowed per case is commonly used in commercial computer-aided detection products. However, the actual scores for each region are not available to users. Therefore, several related issues—such as the effect of this approach on overall performance and on the detection (or the missed detection) of specific masses—have not, to our knowledge, been described in the past.

Our study showed that by limiting the maximum number of allowed regions to be cued in each case, a substantial fraction of false-positive regions can be eliminated with only a small decrease in sensitivity. If one wishes to maintain sensitivity, threshold values can be appropriately adjusted for this purpose. Because most masses were visible on both the craniocaudal and mediolateral oblique mammograms and because the detection performance of computer-aided detection systems is commonly evaluated using case-based sensitivity, our results are quite encouraging. It appears that this approach could reduce the false-positive detection rate of the scheme and possibly eliminate some true-positive region-based detections while retaining the initial (unrestricted number of cues) case-based sensitivity. Although the sensitivity can be maintained using this approach (changing the threshold levels for detection), one does not detect exactly the same true-positive masses. We found that limiting the maximum number of cues allowed per case and adjusting the thresh-

TABLE 2	Performance Levels of Computer-Aided Detection with Constant Sensitivity of 79% as a Function of the Maximum Number of Cued Regions Allowed per Case				
Maximum No. of Cued Regions Allowed per Case	Region-Based Sensitivity[a]		False-Positive Rate[b]		Detection Decision Value of Artificial Neural Network Scores
	No.[c]	%	No.	Per-Image Rate	
No limit	377	66.1	803	0.40	0.565
5	371	65.1	773	0.39	0.560
4	378	66.3	902	0.45	0.500
3	375	65.8	781	0.39	0.470
2	350	61.4	604	0.30	0.360

[a]Detected true-positive cases.

[b]Detected false-positive regions.

[c]Regions.

3. Brem RF, Schoonjans JM. Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. *Clin Radiol* 2001;56:150–154
4. Ciatto S, Turco MR, Risso G, et al. Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography. *Eur J Radiol* 2003;45:135–138
5. Malich A, Azhari T, Bohm T, Fleck M, Kaiser WA. Reproducibility: an important factor determining the quality of computer aided detection (CAD) systems. *Eur J Radiol* 2000;36:170–174
6. Zheng B, Ganott MA, Britton CA, et al. Soft-display mammographic reading with different computer-assisted detection cueing environments: preliminary findings. *Radiology* 2001;221:633–640
7. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computer assisted detection of interval breast cancers. *Eur J Radiol* 2001;39:104–110
8. D'Orsi CJ. Computer-aided detection: there is no free lunch. *Radiology* 2001;221:585–586
9. Malich A, Marx C, Facius M, Boehm T, Fleck M, Kaiser WA. Tumor detection rate of a new commercially available computer-aided detection system. *Eur Radiol* 2001;11:2454–2459
10. Hoffmeister JW, Rogers SK, DeSimio MP, Brem R. Determining efficacy of mammographic CAD systems. *J Digit Imaging* 2002;15:198–200
11. Nishikawa RM, Yarusso LM. Variations in measured performance of CAD schemes due to database composition and scoring protocol. *Proc SPIE* 1998;3338:840–844
12. Kallergi M, Carney GM, Gaviria J. Evaluating the performance of detection algorithms in digital mammography. *Med Phys* 1999;26:267–275
13. Nishikawa RM, Giger ML, Doi K, et al. Effect of case selection on the performance of computer-aided detection schemes. *Med Phys* 1994;21:265–269
14. Zheng B, Chang YH, Gur D. Adaptive computer-aided diagnosis scheme of digitized mammograms. *Acad Radiol* 1996;3:806–814
15. Zheng B, Sumkin JH, Good WF, Maitz GS, Chang YH, Gur D. Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression: an assessment. *Acad Radiol* 2000;7:595–602
16. Zheng B, Chang YH, Gur D. Computerized detection of masses in digitized mammograms using single image segmentation and multilayer topographic feature analysis. *Acad Radiol* 1995; 2:959–966

Fig. 5.—Scatterplot shows sizes and subtlety ratings distributions for 34 masses that were undetected by both case-based and image-based scoring methods. ◆ = no limit to number of regions in each case that may be cued as showing positive findings, ✕ = maximum number of regions that may be cued is ≤ 2.

old appropriately increased computer-aided detection sensitivity in the subset of smaller masses. In general, this effect is desirable in that it could reduce the number of regions that have to be ruled out by the radiologist. We caution that the use of this approach may not yield improvements of similar magnitude in the clinical environment with a substantially different distribution of truly positive and truly negative cases.

It should be noted that the size and subtlety ratings of masses in the data set were somewhat conservative. In Figures 1 and 2, we used the larger of the sizes computed for a mass from the two mammographic views and presented the less subtle rating for the same mass. Hence, distribution based on image or region would show a somewhat smaller average mass size and a more subtle data set.

Only malignant masses were considered true-positive identifications in this study. In visually assessing the false-positive regions with higher scores (e.g., > 0.7), we found that 19% (40/213) of these regions represented well-defined benign masses (i.e., round benign masses with high contrast and relatively sharp margins). Considering the detection of benign masses as either true-positive or false-positive may have a substantial impact on the evaluation of computer-aided detection performance levels. Because of the approach we used to reduce the number of cued regions per case and because of the size and diversity of the data set used, we believe that our results are not unique to our own computer-aided detection scheme.

References

1. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 2000;215:554–562
2. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 2001;220:781–786

Bin Zheng, PhD
Lara A. Hardesty, MD
William R. Poller, MD
Jules H. Sumkin, DO
Sara Golla, MD

Author contributions:
Guarantor of integrity of entire study, B.Z.; study concepts, B.Z.; study design, B.Z., J.H.S.; literature research, B.Z., S.G.; clinical and experimental studies, B.Z., W.R.P.; data acquisition, B.Z., L.A.H., S.G.; data analysis/interpretation, B.Z.; statistical analysis, B.Z.; manuscript preparation, B.Z., S.G., L.A.H.; manuscript definition of intellectual content, B.Z.; manuscript editing, B.Z., S.G., L.A.H.; manuscript revision/review and final version approval, all authors.

Mammography with Computer-aided Detection: Reproducibility Assessment—Initial Experience[1]

PURPOSE: To examine the performance and reproducibility of a commercially available computer-aided detection (CAD) system with a set of mammograms obtained in 100 patients who had undergone biopsy after positive findings at mammography.

MATERIALS AND METHODS: One hundred positive mammographic examinations (four views each), depicting 96 masses and 50 microcalcification clusters, were scanned and analyzed three times by the CAD system. Reproducibility of detection sensitivity and the individual CAD-generated cues in the three images were examined. Both abnormality- and region-based detection sensitivities were compared.

RESULTS: Forty-eight (96.0%) of 50 microcalcification clusters were marked on all three images in the abnormality-based analysis. Of the remaining two clusters, one was marked in two images and one was marked in only one. The abnormality-based sensitivity for mass detection ranged from 66.7% (64 of 96) to 70.8% (68 of 96). The system generated identical patterns (including images with and those without cues) for all three images in 53.3% (213 of 400) of images. For true-positive cluster regions, 88.9% (80 of 90) were marked at the same location in all images. For true-positive mass regions, 69.5% (82 of 118) were marked at the same locations in all images. In false-positive detections, only 44.0% (81 of 184) of false-positive mass regions and 31.9% (38 of 119) of false-positive cluster regions were marked at the same locations on all three images.

CONCLUSION: Reproducibility of marked regions generated by the CAD system is improved from that reported previously, largely as a result of the substantial reduction in the false-positive detection rates. Reproducibility of true-positive identification of masses remains an important issue that may have methodologic and clinical practice implications.

Mammography is a common and effective method with which to screen for early detection of breast cancer, to interpret mammograms, and particularly to identify subtle masses and microcalcification clusters surrounded by complex breast tissue patterns, but it is a difficult and time-consuming task. Findings in studies show that from 10% to 30% of breast cancers that are visible on mammograms during retrospective readings are missed during the original interpretations for various reasons (1–3). One well-documented method to reduce false-negative rates in mammography is the use of an independent double-reading approach (4,5). However, this approach is both inefficient and costly. As a result, after intensive research and substantial improvements in the past 2 decades, computer-aided detection (CAD) systems have been developed to provide radiologists with a "second opinion" when they identify suspicious regions for masses or microcalcification clusters. In the current study, we used one of three commercially available CAD systems that have been approved by the U.S. Food and Drug Administration and are used for this purpose.

Because of the potential importance of CAD systems in the clinical environment, several studies (6–10) have been conducted recently to evaluate the performance of CAD systems

alone and their possible effect on diagnostic performance of radiologists under a variety of clinical conditions. In one recent study involving 12,860 patients in a community breast center, use of CAD resulted in a 19.5% increase in the number of cancers detected without undue effect on the recall rate (from 6.5% to 7.7%) (6). In another large retrospective study, a false-negative rate of 21% was found when 14 radiologists interpreted mammograms, and the CAD system correctly marked 77% of these missed cases (7). Thus, researchers claim that CAD cueing could potentially reduce this false-negative rate by as much as 77% without an increase in the recall rate (8). On the other hand, findings in a different study showed that despite high (and clinically viable) sensitivity, the CAD system had no effect on radiologist performance (including sensitivity and specificity) (9). These researchers suggested that perhaps the many false-positive markings influenced the radiologists not to have sufficient confidence in the CAD results to alter their original interpretations (9). Results in another retrospective study demonstrated that the performance of a CAD system could affect the performance of radiologists in the detection of masses and microcalcification clusters. Highly performing CAD schemes with high sensitivity and a low false-positive rate could improve radiologists' performance significantly, while poorly performing CAD schemes could significantly ($P < .01$) decrease readers' performance (10).

An important issue related to the use of CAD is the reproducibility of results. In one study, an early version of Image-Checker (R2 Technology, Los Altos, Calif) was evaluated, and the authors suggested that its reproducibility may be insufficient for the routine clinical environment (11). Recently, a new version of the software was used, which improves the detection sensitivity and specificity (12). In the version used in the current study (ImageChecker, version 2.0), the stated detection sensitivity for the cancer cases was increased from 83.7% to 90.4% (including an increase in mass detection from 74.7% to 85.7% and an essentially unchanged performance for microcalcification detection of more than 98%). At the same time, the false-positive rate was reduced substantially from approximately 1.0 per image to 0.5 per image (or 4.1–2.06 false-positive cues per four views in true-negative cases) (12). The purpose of our study was to examine the performance and reproducibility of a commercially available CAD system by using a set of mammograms acquired in 100 patients who had undergone biopsy after positive findings at mammography.

MATERIALS AND METHODS

Cases

During the past several years, a large database (>1,000 cases) of digitized images and associated diagnostic results has been established and managed in our laboratory under an approved institutional review board protocol (informed consent was waived). For the purpose of third-party, we asked a staff member not otherwise related to this current investigation to randomly select 100 mammographic cases (four views each) from the biopsy records of our institution during the years 1999–2001. We requested that 25 of the cases depict microcalcification clusters and 75 cases depict masses as a primary detection finding. At least two-thirds of the cases were to be selected from those proven to be malignant. With the exception of these conditions, cases were selected solely by the staff member from the biopsy records. The selection process did not involve a previous review of any of the images. Therefore, there was no preselection (and potentially biasing) process as related to the average tissue density or the subtlety of the abnormalities depicted in the images.

Each case could involve one or more abnormalities (mass, microcalcification cluster, or both). In these 100 cases, 51 depicted only masses (43 depicted one mass and eight depicted two masses), 12 depicted only microcalcification clusters (11 depicted one cluster and one depicted two clusters), and 37 depicted both masses and clusters (one mass and one cluster). There were no cases with more than three abnormalities depicted. The data set involved 96 verified masses and 50 verified microcalcification clusters. Sixty-five of the 96 masses were malignant, and 31 were benign. Thirty-one of the 50 microcalcification clusters were associated with malignancy, and 19 were benign. By examining all source documents (including pathology reports), the locations of all abnormalities were specified by radiologists.

CAD Evaluation

These 400 images were scanned through the CAD system three times within a period of 3 weeks. After digitization and computation, suspicious masses and microcalcification clusters identified by the CAD system were marked on the output paper images by using the standard identification scheme. The CAD system does not outline the entire mass region or individual microcalcifications in a cluster, only a small star or a triangle is superimposed on the image to indicate the presence of a suspicious region for a mass or a cluster, respectively. The boundaries of masses and clusters were identified visually on the images by a researcher (B.Z.), who consulted with radiologists in cases of ambiguity. If the star was located anywhere inside a true-positive mass region in the image, this mass was considered to be identified correctly by the CAD system. Similarly, as long as a triangle was overlapping any of the microcalcification areas, the mark was considered to represent a true-positive detection. Otherwise, the cue was considered to identify a false-positive region. The processing of each case resulted in three sets of output images.

Data Analysis

The sensitivity, false-positive rate, and reproducibility of the CAD system with these 100 cases (or 400 images) were analyzed for abnormality- and region-based values. In the abnormality-based analysis, the sensitivity is assessed on the basis of the correct marking of at least one true-positive region in either view (craniocaudal, mediolateral oblique, or both), which included 96 masses (65 malignant) and 50 calcifications (31 malignant) in the 100 cases. In cases with more than one abnormality, each was considered to be independent of the others. In the region-based analysis, the abnormality depicted in each view (either craniocaudal or mediolateral oblique) was considered an independent true-positive finding. Sensitivity was then computed on the basis of the number of correctly detected true-positive regions (rather than abnormalities). This approach included 292 positive findings—namely, 96 masses and 50 clusters, each visible on two views. To compare the differences in proportions of correctly detected abnormalities among replicated images, the pairwise McNemar test was applied to the data set.

RESULTS

Tables 1 and 2 summarize the performance of the CAD system with respect to mass and microcalcification cluster detection in each of the three scans. Abnormality-based sensitivity for mass detection ranged from 66.7% (64 of 96) to 70.8% (68 of 96). Although scan 2 yielded highest sensitivity for mass detection (68 of 96), scan 1 depicted the highest number of malignant masses (47 of 65). For microcalcification cluster detection, 48 of 50 clusters were

TABLE 1
Mass Detection Performance of CAD System during Each Scan

Scan No.	Sensitivity (all cases)		Sensitivity (malignant cases only)		False-Positive Rate per Image
	Abnormality Based (%)	Region Based (%)	Abnormality Based (%)	Region Based (%)	
1	69.8 (67 of 96)	52.1 (100 of 192)	72.3 (47 of 65)	54.6 (71 of 130)	0.33 (130 of 400)
2	70.8 (68 of 96)	52.6 (101 of 192)	70.8 (46 of 65)	52.3 (68 of 130)	0.33 (131 of 400)
3	66.7 (64 of 96)	51.0 (98 of 192)	69.2 (45 of 65)	51.5 (67 of 130)	0.31 (125 of 400)

TABLE 2
Microcalcification Cluster Detection Performance of CAD System during Each Scan

Scan No.	Sensitivity (all cases)		Sensitivity (malignant cases only)		False-Positive Rate per Image
	Abnormality Based (%)	Region Based (%)	Abnormality Based (%)	Region Based (%)	
1	96.0 (48 of 50)	85.0 (85 of 100)	93.5 (29 of 31)	85.5 (53 of 62)	0.17 (69 of 400)
2	98.0 (49 of 50)	87.0 (87 of 100)	96.8 (30 of 31)	87.1 (54 of 62)	0.19 (77 of 400)
3	100 (50 of 50)	86.0 (86 of 100)	100 (31 of 31)	87.1 (54 of 62)	0.20 (79 of 400)

TABLE 3
Number of Times a Mass (or a Region) was Detected

No. of Times Detected	True-Positive Masses	Malignant Masses	True-Positive Mass Regions	Malignant Mass Regions	False-Positive Mass Regions	Total Marked Mass Regions
3 (%)	58 (77.3)	41 (78.8)	82 (69.5)	58 (71.6)	81 (44.0)	163 (54.0)
2 (%)	8 (10.7)	4 (7.7)	17 (14.4)	8 (9.9)	40 (21.7)	57 (18.9)
1 (%)	9 (12.0)	7 (13.5)	19 (16.1)	14 (17.3)	63 (34.3)	82 (27.1)
Total	75	52	118	81	184	302

TABLE 4
Number of Times a Microcalcification Cluster (or a Region) was Detected

No. of Times Detected	True-Positive Clusters	Malignant Clusters	True-Positive Cluster Regions	Malignant Cluster Regions	False-Positive Cluster Regions	Total Marked Cluster Regions
3 (%)	48 (96.0)	29 (93.5)	80 (88.9)	50 (89.3)	38 (31.9)	118 (56.5)
2 (%)	1 (2.0)	1 (3.2)	8 (8.9)	5 (8.9)	30 (25.2)	38 (18.2)
1 (%)	1 (2.0)	1 (3.2)	2 (2.2)	1 (1.8)	51 (42.9)	53 (25.3)
Total	50	31	90	56	119	209

detected by the CAD system on all three images. Two malignant clusters were missed in two of three scans (scans 1 and 2), and one of these clusters was missed in scan 2. With the pairwise McNemar test, no significant ($P > .3$) differences were found in the detection results between any pair of the three scans.

For region-based sensitivity, mass detection ranged from 51.0% (98 of 192) to 52.6% (101 of 192). The total number of masses detected ranged from 98 to 101 in each of the three scans. However, the actual difference in the individual mass regions detected was larger. For example, scan 1 depicted 100 regions and scan 2 depicted 101 regions. However, only 88 of these regions were detected in both images. For the detection of microcalcification clusters, the region-based sensitivity ranged from 85.0% (85 of 100) to 87.0% (87 of 100) for individual cluster regions and from 85.5% (53 of 62) to 87.1% (54 of 62) for malignant clusters.

Although Tables 1 and 2 show that the total number of regions detected in this set of images is relatively constant with all three scans, the locations of the regions detected (in particular, false-positive regions) could differ from scan to scan. In 213 of 400 images, the output results for all three scans were identical, which represents an overall reproducibility of 53.3%. Among these images, 37.6% (80 of 213) had no cues (including neither true-positive nor false-positive cues) in all three scans. For the remaining 320 images, the CAD system marked 511 regions (1.6 cues per image) in three scans (including true-positive cues). Of these 511 marked regions, 281 were identified on all three scans (55% region-based reproducibility).

Tables 3 and 4 summarize the number of true-positive and false-positive masses and microcalcification clusters (including both abnormalities and regions) that were identified in all three scans, two scans, or only one scan. The results show that the reproducibility for the true-positive regions (those identified in all three scans) is substantially higher than that for the false-positive regions. For the true-positive mass regions, the CAD system generated 118 cues in three scans, and 82 (69.5%) of them were marked at the same locations. For the true-positive cued cluster regions, 88.9% (80 of 90) of cues were in the same locations for all three scans. On the other hand, the reproducibility of the false-positive cues was much lower, with a higher fraction of different cues being generated in each scan. Only 44.0% (81 of 184) of the false-positive mass regions and 31.9% (38 of 119) of the false-positive microcalcification cluster regions were marked at the same locations in all three scans.

DISCUSSION

In a previous study, 38.5% (77 of 200) of images had CAD cues that were located congruently in all three scans (11). In the current study, the CAD system generated identical results on 53.0% (213 of 400) of the images. The improvement in reproducibility may be largely a result of the substantial decrease in the false-positive detection rate (from approximately 1.0 to 0.5 per image) (12). When we exclude 80 images that had no CAD cues, the reproducibility in the remaining 320 images was reduced to 41.6% (133 of 320). However, the reproducibility in detecting specific true-positive masses and microcalcification clusters is perhaps more important than the more general image-based reproducibility. It is generally difficult to directly compare the detection performance in two experiments, because different image databases were used and the results depend heavily on the difficulty of the selected cases (13). However, some comparative information can be ascertained. In a previous report, the CAD system performed better for mass detection (86.9% abnormality-based sensitivity) than for microcalcification cluster detection (76.6%) (11), while in the current study, sensitivity for the detection of microcalcification clusters was higher than 96%, and the sensitivity for mass detection was in the range of 70%. These results may indicate that the microcalcification clusters depicted in our data set were easier to detect, and masses depicted in our database were more subtle. The case selection protocol we used should have reduced biases; however, the results presented herein with a small database may not represent the actual performance of the system in a clinical setting. Findings in the current study demonstrated clearly that the issue of reproducibility of image-based CAD systems needed to be investigated further.

It should be noted that we obtained somewhat different results in absolute terms for the benign and malignant cases, but the pattern for the two groups remained similar. All cases in our study were sufficiently suspicious to ultimately warrant a recommendation for biopsy. We believe that at this stage, CAD schemes should be designed and optimized to identify this group of cases, including those that ultimately prove to be benign. It is well known that repeated scanning of the same image results in a slightly different digital value matrix for a variety of technical reasons. In current

CAD systems, a binary threshold is typically used to generate detection marks. Each marked region has a computed score that is above a predetermined threshold; hence, lesions with computed scores that are near the threshold are vulnerable to small changes and may be detected in one image and missed in another. Findings in the present study show that the reproducibility of false-positive cues was much lower than that of true-positive cues (Tables 3 and 4), because the detection scores may be close to the threshold. We did not perform a complete long-term follow-up to confirm that all false-positive cues actually represented negative regions. Should any false-positive detection prove to be a true abnormality, the computed reproducibility level would be lower than that reported herein.

Note that the databases used in this and a previous (11) study were small; hence, the results may not represent the actual reproducibility of CAD systems in the screening environment. Despite this limitation, findings in the two studies highlight an important finding. Current CAD schemes are sensitive to small variations in the digital value matrices that result from repeated scanning of the same images. This may have methodologic and clinical practice implications that need to be addressed. The fact that all abnormalities depicted in the present study were visible on both views indicates that the cases were not particularly subtle and that the findings we report herein, including possible implications, may be magnified in cases that are more difficult to identify visually or when the abnormality is visible only on one view. We suspect that this sensitivity to minor changes in the matrices is not unique to the CAD system evaluated in the current study. Full-field digital mammography systems are rapidly becoming available (14,15). By definition, once an image is acquired, the CAD detection result will be 100% reproducible when the same CAD scheme is applied repeatedly to such an image. To be optimal, however, current CAD schemes may have to be reengineered and reoptimized by using digitally acquired images before these schemes can be applied optimally to full-field digital mammography systems. An investigation on possible effects of repeated image acquisition of the same breast on CAD results is beyond the scope of the present study.

Findings in our preliminary study suggest that sensitivity for the detection of microcalcification clusters is high; as a

result, reproducibility is also high. These results are achieved at a low false-positive detection rate; hence, it is a useful tool during the diagnostic process. Our results raise the important question about the possible need to maintain records of CAD cues as available during the interpretation of the individual cases. This may become an even more important issue as cancer detection continues to progress toward an earlier stage (hence, a more subtle appearance) on the average. Detailed documentation of all available information at the time of diagnosis is not always done, particularly since information is often provided verbally. In the case of screening mammographic interpretation, however, the presence of a malignancy that was visible (in retrospect) on a previous mammogram and in which a follow-up scan of the original images in a CAD system may produce a true-positive identification, could present a medicolegal problem. It will be difficult to argue that the abnormality in question was not identified as suspicious on the original image. Findings in our preliminary study suggest that this may be the case in a noticeable fraction of mass cases (approximately 20%, as shown in Table 3).

The current practice associated with the use of CAD in the mammographic environment is not clear on whether a record of the CAD results used during the case interpretation should be retained. Until mass detection is substantially improved, results in our study suggest that such a practice should be considered. Interestingly, although largely impractical, our study findings clearly suggest that at this level of performance, multiple repeated scans of each case could be acquired to improve the performance of CAD schemes.

References

1. Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. Radiology 1992; 184:613–617.
2. Harvey JA, Fajardo LL, Innis CA. Previous mammograms in patients with impalpable breast carcinomas: retrospective vs blinded interpretation. AJR Am J Roentgenol 1993; 161:1167–1172.
3. Goergen SK, Evans J, Cohen GP, MacMillan JH. Characteristics of breast carcinomas missed by screening radiologists. Radiology 1997; 204:131–135.
4. Thurfjell EL, Lernevall KA, Taube AAS. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994; 191:241–244.
5. Hendee WR, Beam C, Hendrick E. Proposition: all mammograms should be double-read. Med Phys 1999; 26:115–118.
6. Freer TW, Ulissey MJ. Screening mammography with computer-aided detec-

tion: prospective study of 12,860 patients in a community breast center. Radiology 2001; 220:781–786.

7. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215:554–562.

8. Birdwell RL, Ikeda DM, O'Shauhhnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001; 219:192–202.

9. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computed assisted detection of interval breast cancers. Eur Radiol 2001; 39:104–110.

10. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic reading with different computer-assisted detection cueing environments: preliminary findings. Radiology 2001; 221:633–640.

11. Malich A, Azhari T, Bohm T, Fleck M, Kaiser WA. Reproducibility: an important factor determining the quality of computer aided detection (CAD) systems. Eur Radiol 2000; 36:170–174.

12. Castellino RA, Roehrig JR, Zhang W. Improved computer-aided detection (CAD) algorithms for screening mammograms (abstr). Radiology 2000; 217(P): 400.

13. Nishikawa RM, Giger ML, Doi K, et al. Effect of case selection on the performance of computer-aided detection schemes. Med Phys 1994; 21:265–269.

14. Lewin JM, Hendric RE, D'Orsi CJ, et al. Comparison of full-field digital mammography with screen-film mammography for caner detection: results of 4,945 paired examinations. Radiology 2001; 218:873–880.

15. Venta LA, Hendrick RE, Adler YT, et al. Rates and causes of disagreement in interpretation of full-field digital mammography and screen-film mammography in a diagnostic setting. AJR Am J Roentgenol 2001; 176:1241–1248.

Integrated density of a lesion: A quantitative, mammographically derived, invariable measure

Yuan-Hsiang Chang,[a)] Walter F. Good, Joseph K. Leader, Xiao-Hui Wang, Bin Zheng,
Lara A. Hardesty, Christiane M. Hakim, and David Gur
*From the Department of Radiology, University of Pittsburgh and Magee-Womens Hospital, Pittsburgh,
Pennsylvania 15213*

A method for quantitatively estimating lesion "size" from mammographic images was developed and evaluated. The main idea behind the measure, termed "integrated density" (ID), is that the total x-ray attenuation attributable to an object is theoretically invariant with respect to the projected view and object deformation. Because it is possible to estimate x-ray attenuation of a lesion from relative film densities, after appropriate corrections for background, the invariant property of the measure is expected to result in an objective method for evaluating the "sizes" of breast lesions. ID was calculated as the integral of the estimated image density attributable to a lesion, relative to surrounding background, over the area of the lesion and after corrections for the nonlinearity of the film characteristic curve. This effectively provides a measure proportional to lesion volume. We computed ID and more traditional measures of size (such as "mass diameter" and "effective size") for 100 pairs of ipsilateral mammographic views, each containing a lesion that was relatively visible in both views. The correlation between values calculated for each measure from corresponding pairs of ipsilateral views were computed and compared. All three size-related measures (mass diameter, effective size, and ID) exhibited reasonable linear relationship between paired views ($r^2 > 0.7, P < 0.001$). Specifically, the ID measures for the 100 masses were found to be highly correlated ($r^2 = 0.9, P < 0.001$) between ipsilateral views of the same mass. The correlation increased substantially ($r^2 = 0.95$), when a measure with linear dimensions of length was defined as the cube root of ID. There is a high degree of correlation between ID-based measures obtained from different views of the same mass. ID-based measures showed a higher degree of invariance than mass diameter or effective size. © *2003 American Association of Physicists in Medicine.*
[DOI: 10.1118/1.1582571]

I. INTRODUCTION

Mammography has been frequently recommended as a routine screening tool to detect breast cancers at an earlier stage.[1] While it has been shown that mammographic screening can substantially reduce cancer-related mortality and morbidity,[2,3] identifying breast cancers in the screening environment with high sensitivity and specificity is a difficult task due to the low expected cancers detected in a large volume of cases and the complex patterns as depicted on mammograms.[4,5] Detection and diagnostic accuracy of breast cancers using mammograms can be improved using quantitative analysis of masses.[6-9] Studies of cases with prior examinations demonstrated that a change in the density or contour of a mass over time can be an indicative sign of malignancy.[10] Other studies showed that the change in mass size was one of the dominant factors in determining breast cancer prognosis.[11,12] However, inter- and intra-observer variability, when visually and subjectively describing a mass or its change over time between consecutive examinations, makes this assessment quite difficult and often unreliable.[13,14]

Quantitative measurements and analyses of masses have been widely used in computer-aided detection and character-ization (discrimination) schemes.[15,16] A large number of features (including morphologic, texture, and density based[17,18]) have been investigated in an attempt to quantitatively and objectively represent masses. Currently, there is no standard that defines a mass and its surrounding background. Without such a standard, studies have shown that measured contrast values of a mass region could change substantially if the definition of the surrounding background of a mass was varied.[19] Therefore, measurements of mass contrast and other related features are frequently scheme dependent. In addition, due to the wide variation of tissue presentation resulting from breast compression, image projection, and field nonuniformity, a large number of computed image-based features of a mass, as measured from different images, is not a constant.[20,21] Therefore, it would be desirable to define a measure of a mass that is invariant to tissue deformation and the projection view. It has been theoretically shown that under a few conditions, the total attenuation (termed here integrated density or ID) of an object is an invariant quantity with respect to geometrical deformation of an object in two-dimensional projected images (such as mammograms).[22] There is no experimental validation that in fact ID is an invariant measure.[22] In this study, we experimentally evalu-

ate integrated density (ID) as an invariant measure representing masses depicted on mammograms. After correcting for film nonlinearity, and estimating the density of underlying tissue, we assess whether ID as computed for the same 100 masses depicted during the same examination on both CC and MLO views is invariant. The purpose of this preliminary study is intended only to assess the degree of ID invariance with respect to breast compression and projection views, and not changes over time.

II. MATERIALS AND METHODS

A. Defining integrated density

Assuming a monoenergetic x-ray source, for a mass present in a breast, the quantity

$$\int\int\int_V \mu_M \, dV,$$

where V is the volume occupied by the mass and μ_M is its local x-ray attenuation coefficient, is invariant, but it cannot be measured directly from the image. However, if it is assumed that the attenuation coefficient of the breast tissue surrounding the mass (μ_N) can be estimated with reasonable accuracy, then the quantity

$$\int\int\int_V (\mu_M - \mu_N) dV = \int\int\int_V \Delta\mu \, dV,$$

remains essentially invariant and can be approximated from an image.

To demonstrate this, consider the density of a single pixel within the image projection of the mass. In the case of a mammographic film image, for which the digitized density values have been adjusted for the film's characteristic curve, the corrected density at an image pixel, $D(u,v)$, can be calculated as

$$D(u,v) = \text{Const} + \gamma \log\left[E \cdot \exp\left(- \int_L \mu_N \, dx \right. \right.$$
$$\left. \left. - \int_{L_M} (\mu_M - \mu_N) dx \right) \right]$$
$$= \text{Const} + \gamma \log E - \gamma \int_L \mu_N \, dx$$
$$- \gamma \int_{L_M} (\mu_M - \mu_N) dx,$$

where E is the overall exposure, γ is the film's gamma, the first integral is taken over the whole breast tissue (L) as projected onto the image, and the second integral is taken over the mass region (L_M) as projected onto the image. The first two terms on the left of the expanded expression correspond to background density, D_{BKG}, of the film (i.e., film base plus nonattenuated exposure including scatter contribution). The third term is the reduction in density resulting from sur-

rounding tissues, D_{NORMAL}, and the final term is the reduction in density attributed to the presence of the mass. This pixel density can be rewritten as

$$D(u,v) = D_{\text{BKG}} - D_{\text{NORMAL}} - \gamma \int_{L_M} \Delta\mu \, dx,$$

or, rearranging this equation,

$$\gamma \int_{L_M} \Delta\mu \, dx = D_{\text{BKG}} - D_{\text{NORMAL}} - D(u,v).$$

If we integrate this over the region, R, defined by the projection of the mass, we obtain

$$\int\int_R \left(\gamma \int_{L_M} \Delta\mu \, dx \right) du \, dv$$
$$= \int\int_R (D_{\text{BKG}} - D_{\text{NORMAL}}) du \, dv - \int\int_R D(u,v) du \, dv,$$

and simplifying the left-hand side,

$$\gamma \int\int\int_V \Delta\mu \, dV = \int\int_R (D_{\text{BKG}} - D_{\text{NORMAL}}) du \, dv$$
$$- \int\int_R D(u,v) du \, dv.$$

The left-hand side of this expression is the film's γ times a quantity that is expected to be essentially invariant; hence, it should be invariant for a particular type of film (in the "linear" range). Thus, we define ID as

$$\text{ID} = \gamma \int\int\int_V \Delta\mu \, dV,$$

which can be approximated as

$$\text{ID} \approx (D_{\text{BKG}} - D_{\text{NORMAL}} - D_{\text{MASS}}) A_R = C \cdot A_R.$$

Therefore, ID is represented by the area of the mass, A_R, multiplied by the average contrast difference between the mass and the surrounding tissue, where the mass area is defined to be the projected area associated with the interior of the mass and the average mass contrast, C, is defined as the average difference in linearized densities between the underlying tissue background and the mass itself.[19] It should be emphasized that x-ray beam hardening is ignored in this simplification and scatter radiation is assumed to contribute a relatively uniform distribution in both the mass and background areas. Hence, it can be represented as a constant in the background term. Determining the relative change in log-exposure, from film density in mammograms, involves approximating the density due to the combination of background (including scatter) and normal tissues (i.e., $D_{\text{BKG}} - D_{\text{NORMAL}}$), which would be present if the lesion did not exist. This can be estimated as the mean pixel value of the area outside the lesion. It should be noted that such estima-

tion is based on the assumption that the "normal" tissue is distributed "uniformly" over the area of a projected lesion and the surrounding area. Therefore, measurement of the surrounding area would closely resemble the measurement of the underlying tissue if the lesion was absent. Because of the assumptions mentioned above, it was the intent of this study to assess the extent to which ID as simply computed from the image remains invariant to deformation or projection view.

Film characteristic curve linearization: Digitized film density values were corrected for the characteristic curve to produce values that were linearly proportional to log-exposure (in the linear range). First, the laser film digitizer was routinely calibrated to assure that film optical density (OD) was linearly translated into digitized pixel values in the density range of interest. Second, film OD values were corrected so that they were linearly proportional to log-exposure units. A generic curve was used for this purpose using the data for the specific mammography film used in our facility (Eastman Kodak Min-R 2000 film).[23] For computational ease, the generic curve was represented by a spline function.[23] To define ID in terms of more familiar "density" units, rather than log-exposure units, we converted back exposure values to linearized density values by fitting a straight line to this curve. As a result, for each OD value, we computed the corresponding log-exposure unit using the spline function and then converted it to a "linearized" OD value using the "fitted" line.

Delineating mass regions: For each mass, the corresponding pair of mammographic views was reviewed by experienced mammographers who initially identified a central point in the projection of the mass as depicted in two projections. A semi-automatic delineation of mass boundaries was then performed using a region growing routine similar to the technique described by Matsumoto *et al.*[24] For each projection, the location representing the central point (pixel) of the mass was first identified on digitized mammograms. Based on the initial location, the algorithm automatically determined a "transition layer," where a substantial change in region growth and margin irregularity was observed. All pixels within the identified region boundary were considered to be in the region (R) depicting the mass. Mass delineation could affect the results since both the area of the mass (A_R) and the average density within the mass (D_{MASS}) are used to compute ID.

Measuring mass background: Once a mass region (R) had been delineated, the geometric center-of-mass of the lesion was computed, and the maximum distance of this point to the lesion boundary was determined. For background density estimation, a region was defined as all tissue regions outside the lesion, within a circle centered at the center of mass (R'), with a radius difference (Δ radius) longer than the maximum distance to the lesion's boundary. Linearized density values within the background region (R') were averaged.

To estimate the density of normal tissue in the areas where the mass overlaps in the projection image, we used the average density value of the surrounding area outside the mass. The radius difference was initially chosen as Δ radius
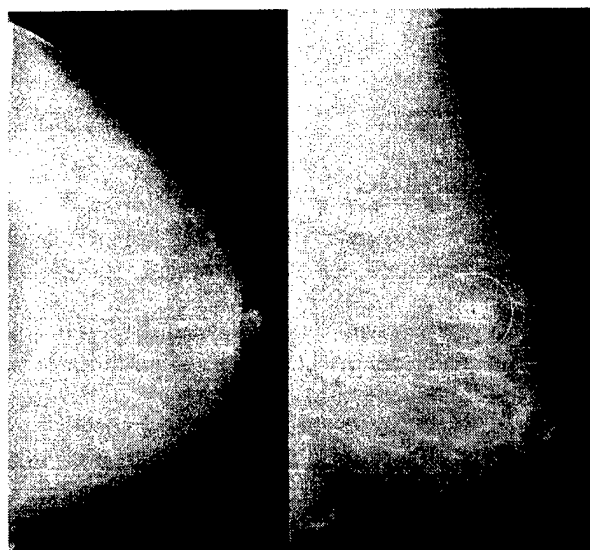

FIG. 1. An example of a pair of images depicting a mass in two views. The mass is relatively visible in both the CC (left) and MLO (right) views. Superimposed on each view are the mass center ("+"), the computed mass boundary, and the corresponding mass background (circular region excluding the interior of the mass region).

$=1$ cm for the background estimation. However, since the definition of the background region can affect the computed ID, we investigated this issue in the following manner. Based on the distribution of mass diameters in our database, we selected four values for the radius differences (Δ radius $=5$ mm, 7.5 mm, 10 mm, or 12.5 mm) to be considered as the background area. IDs were estimated from the CC and MLO views using the different background definitions, and the correlations between computed ID values for the corresponding views were computed.

ID computation for each mass: The area of the mass, A_R, was obtained by counting the number of pixels inside the region R and converting it to an area (one pixel represented an area of 100×100 μm^2 or 0.01 mm^2). Average mass contrast, C, was computed as the difference in average density values within the regions R and R'. Finally, ID is the product of the mass area and the average mass contrast (i.e., ID$\approx C \cdot A_R$). This process was performed independently for each mammographic view.

Figure 1 demonstrates an example of the regions analyzed in one case. The mass was clearly visible in both the CC view and MLO views. In each view, an irregular boundary marks the mass region as automatically determined by the software. The background region of surrounding tissue that was used in the analysis is also shown.

B. Other relevant measures

In addition to the computed ID, two mammographically based measures were derived for each projection of a mass, as follows: (1) Mass diameter was defined as the maximum diameter (or longest axis) of the mass as depicted in the image (in mm); and (2) effective size was defined as the squared root of the product of maximum and minimum di-

ameters (longest axis and shortest axis) and was also measured in mm.[25] These measures were used for comparison with ID, as these are frequently used in the clinical environment and in CAD assessments.

C. Dataset

A total of 100 verified cases were selected from our database of patients who have undergone screening mammography in one of our clinical facilities. Each case included an ipsilateral pair of craniocaudal (CC) and mediolateral oblique (MLO) mammographic images. CC and MLO views of the same breast from the same examination were used. Cases were selected only if a well-defined mass was depicted on both views. Of the 100 selected masses, 64 were malignant and 36 benign. All films were digitized using a laser film digitizer (Lumisys, Eastman Kodak Co., Rochester, NY) at $100 \times 100~\mu m^2$ pixel size and 12-bit contrast resolution. The laser film digitizer was routinely calibrated to assure that film optical density (OD) is linearly translated into digital pixel values in the range of 0.2 to 3.8 OD (1 pixel value = 0.001 OD). The three measures (ID, diameter, and effective size) were estimated for both views of each of the 100 masses in the dataset.

D. Evaluation

In this study we compared the results of the three measures as computed independently for each of the two views. For each measure, we computed the correlation (Pearson's r^2) between values computed from the CC views and those from the MLO views. The results for each measure were also plotted in corresponding scattergrams.

Because of the units associated with these measures, it is suboptimal to compare directly the correlations for ID with those computed for the other two measures. ID is proportional to volume, while the other two measures are defined in units of lengths. Therefore, we defined a cube root of ID as a more appropriate measure for comparisons (which is given in unit of length) and report the results of this measure, as well.

To evaluate size-dependent differences between paired measurements obtained from the two views for each of the 100 masses, we divided the database into three subsets (<10 mm, 10–20 mm, and >20 mm). The absolute values of the differences and the range were evaluated for each of the subsets.

Geometric eccentricity of a mass is one factor that could affect the results. Therefore, we categorized masses into two groups by their eccentricity, and repeated the analyses for each of the subgroups. We classified cases into the subgroups using the two diameters d_{CC} and d_{MLO}, obtained from the two views, and computed a ratio $\hat{e}$ for each mass as: $\hat{e} = \max(d_{CC}/d_{MLO}, d_{MLO}/d_{CC})$. All masses for which $\hat{e} \leq 1.1$, were assigned to one group (more "spherical") and the remaining masses (more "eccentric") were assigned to the other group. Masses with substantially different diameters as depicted on the CC and MLO views typically exhibit
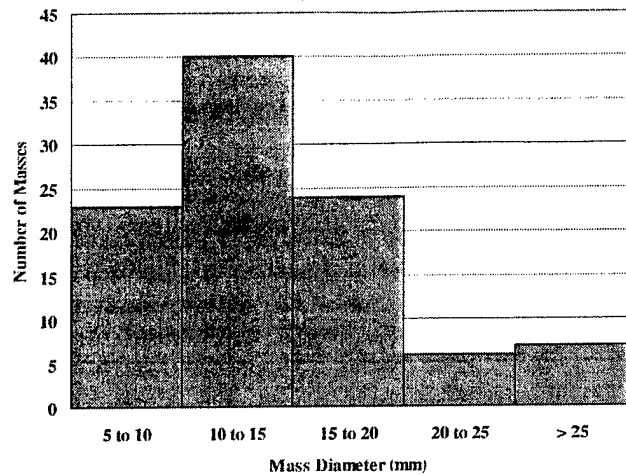

FIG. 2. Distribution of the mass sizes for the 100 masses used in this study. Mass size was determined by averaging maximum diameters (mm) obtained from the two ipsilateral views (CC and MLO).

high eccentricity. Thus, a relatively low correlation for diameter or effective size was expected in this group.

III. RESULTS

Figure 2 shows the distribution of mass sizes. The size of each mass was determined by averaging the two maximum diameters obtained from the two views (CC and MLO). As can be seen from the figure, the 100 selected masses ranged in sizes from relatively small (<10 mm) to quite large (>25 mm).

Table I shows the linear correlation coefficients (Pearson's r^2) between matched pairs of measures (i.e., mass diameter, effective size, mass area, mass contrast, and ID) from the two views. Figure 3 shows the corresponding scattergrams for (a) mass diameter, (b) effective size, (c) ID, and (d) cube root of ID. The size-related measures (mass diameter and effective size) were found to follow a reasonable linear relationship ($P < 0.001$). Despite the relatively weaker correlation exhibited by the mass contrast ($r^2 = 0.45$), the ID measure, which is the product of mass area and relative contrast, highly correlated between paired measurements ($r^2 = 0.9, P < 0.001$). Figure 3 also demonstrates the scattergram of the cube root of ID ($\sqrt[3]{mm^2 \cdot \Delta OD}$), which allows for a more valid comparison with the mass diameter and effective size

TABLE I. Linear correlation coefficients (*Pearson's* r^2) of various mass size-related measures (mass diameter, effective size, mass area, mass contrast, and ID) as measured from the CC and MLO view using 100 masses for the assessment.

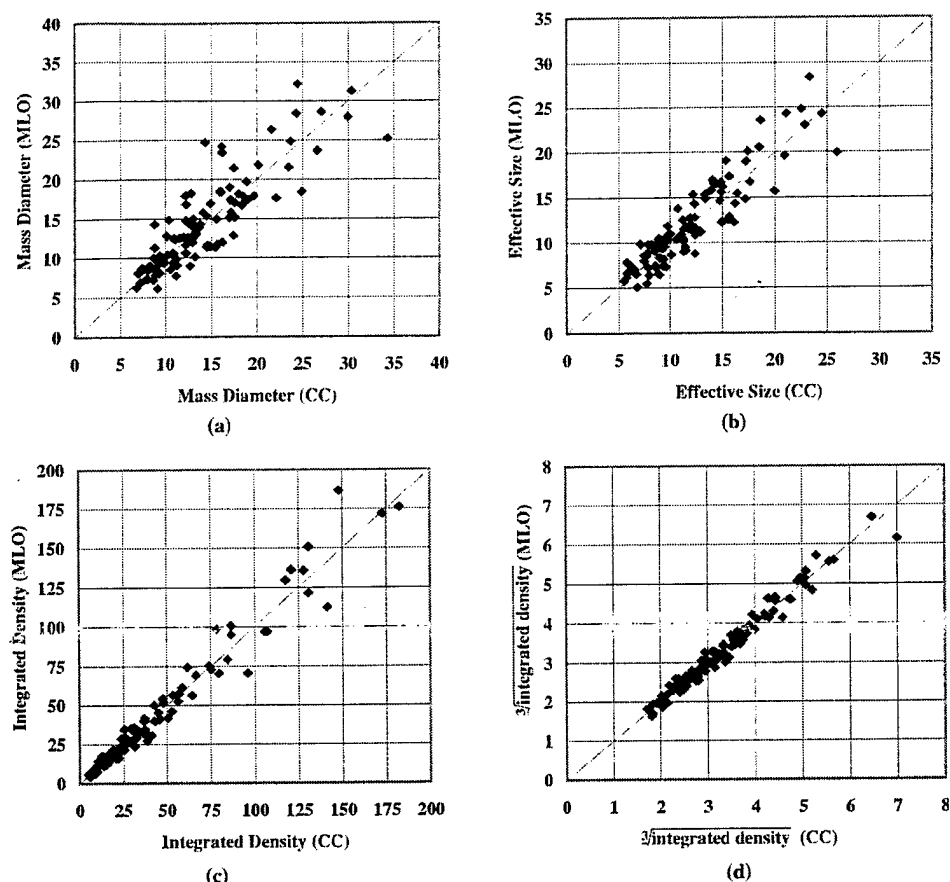|	Mass diameter	Effective size	Mass area	Mass contrast	ID
Correlation coefficient r^2	0.74	0.79	0.83	0.45	0.90

Fig. 3. Scattergrams of the quantitative measures (a) mass diameter, (b) effective size, (c) ID, and (d) cube root of ID, as computed from paired ipsilateral views of 100 masses. The diagonal line represents identical measures.

since it is represented by a similar dimension. The correlation coefficient for the data shown for this measure is substantially higher $r^2 = 0.95$.

Table II shows the ranges of the absolute differences between paired measures obtained from the two views for the different subsets of mass sizes. Three subsets (mass diameter $\leqslant 10$ mm, 10–20 mm, and >20 mm) are

shown. Differences for all measures increase as the mass diameter increase. Measured eccentricity values were 1.14 ± 0.12, 1.17 ± 0.16, and 1.19 ± 0.15 for the three groups, respectively.

Table III shows the linear correlation coefficients of the measures, for each of the subgroups of masses based on eccentricity. As can be seen from this table, both ID and cube root of ID correlated well for the different types of masses, those with low or high eccentricity. The other two measures

TABLE II. Absolute values of differences between paired measures obtained from two corresponding views for the three subsets of masses segmented by mass size.

	Mass diameter		
	<10 mm	10–20 mm	>20 mm
Number of masses	23	64	13
Difference in mass diameter Δmm	1.1 ± 0.8	2.2 ± 1.9	4.0 ± 2.9
Difference in effective size Δmm	1.1 ± 0.7	1.6 ± 1.0	2.6 ± 1.8
Difference in ID $\Delta(mm^2 \cdot \Delta OD)$	2.2 ± 1.5	5.1 ± 5.7	20.2 ± 26.9
Difference in cube root of ID $\Delta(\sqrt[3]{mm^2 \cdot \Delta OD})$	0.13 ± 0.1	0.14 ± 0.07	0.21 ± 0.22

TABLE III. Linear correlation coefficients (*Pearson's* r^2) between corresponding paired views of the same masses for the two groups of cases segmented by eccentricity.

	Eccentricity ($\hat{e}$)	
	$\leqslant 1.1$	>1.1
Number of masses	44	56
Mass diameter (r^2)	0.98	0.58
Effective size (r^2)	0.85	0.74
ID (r^2)	0.92	0.90
Cube root of ID (r^2)	0.95	0.95

TABLE IV. Linear correlation coefficients (*Pearson's* r^2) between corresponding paired views for the malignant and benign masses.

	Malignant masses	Benign masses
Number of masses	64	36
ID (r^2)	0.90	0.92

exhibited lower correlations for the subset of masses with high eccentricity ($\hat{e}$).

Table IV shows the correlation of ID values between CC and MLO views for malignant (64) and benign (36) cases, respectively. While malignant masses generally exhibit more irregular boundary, our experimental results showed in this group of well-defined masses that the ID correlations were similar (0.90 and 0.92 for malignant and benign masses, respectively).

Table V shows the range of averaged ID values and correlation coefficients between paired CC and MLO views when different areas were used for background definition (i.e., Δ radius = 5 mm, 7.5 mm, 10 mm, or 12.5 mm, respectively). Although the ranges of ID measurements varied with respect to background definitions, ID values were well correlated between the paired views in all measurements.

IV. DISCUSSION

We evaluated a quantitative measure for estimating the "sizes" of masses as depicted in mammograms. For 100 verified masses, ID was "estimated" as the product of its area and average linearized contrast. ID was found to be relatively "invariant" between the two views (CC and MLO). It was shown to be a better measure than the others tested in this study, and in particular it is better than the other measures for the subset of masses that are more eccentric. When ID was transformed to a measure with a unit comparable to length (by taking the cube root), its performance increased substantially, resulting in a correlation coefficient of $r^2 = 0.95$.

Computed ID values are dependent on the segmentation method and the definition of the surrounding background region used. Therefore, ID remains scheme dependent. However, despite a reasonably wide distribution of mass sizes and shapes and the background areas evaluated, our results suggest that ID is reasonably invariant with respect to the image projection (view).

TABLE V. Range of averaged measured ID values and the correlation coefficients between paired views for different background definitions. Δ radius = 5 mm, 7.5 mm, 10 mm, and 12.5 mm were used.

	Radius difference (Δ radius) for background definition			
	5 mm	7.5 mm	10 mm	12.5 mm
Range of averaged IDs ($mm^2 \cdot \Delta OD$)	40.7±43.7	44.5±47.6	48±51.6	51.2±55.7
Correlation (*Pearson's* r^2)	0.92	0.91	0.90	0.88

Because of the presence of the mass, one can only approximate the measurement from the surrounding area using the assumption that the normal tissue in the area of the projected mass and surrounding area are the same. This assumption is but one source of error in the computed ID value that could affect its invariance properties. This preliminary study was intended mainly to assess the performance of ID as an invariant measure of mass size. Hence, we included only cases with relatively well-defined masses that were clearly visible on both views. As a result, the reported correlation coefficients for all measures applied to our dataset are likely to overstate the performance of such a measure in a dataset that includes more subtle or somewhat ill-defined cases.

The assumptions that the measure will not be significantly affected by scatter, film gamma, and exposure factors (e.g., geometry, kVp, filtration, and field uniformity[21]) seem to be reasonable for this purpose, but the desired characteristic (invariance) will have to be experimentally confirmed for different experimental conditions.

Our findings suggest that it may be possible to achieve a relatively reproducible measure for a given mass over a rather wide range of conditions and different mammographic views. We appreciate the fact that the definition of some of the measurements of interest in this work were simplified and could be more accurately described. However, the intent was to develop a relatively easy measure to compute, perhaps at the cost of being somewhat less rigorous and precise.

V. CONCLUSION

We have developed a method for deriving a quantitative measure of lesion "size," termed integrated density or ID, that was found to be reasonably invariant between paired projection views of the same breast. Our experimental results in a set of 100 well-defined masses demonstrated a high degree of correlation between ID-based measures obtained from ipsilateral paired views of the same breast. ID-based measures showed a higher degree of correlation when compared with other traditional size-related measures, such as mass diameter or effective size.

[a])Electronic mail: ychang@mail.magee.edu
[1]S. A. Feig, C. J. D'Orsi, R. E. Hendrick, V. P. Jackson, D. B. Kopans, B. Monsees, E. A. Sickles, C. B. Stelling, M. Zinninger, and P. Wilcox-Burchalla, "American College of Radiology guidelines for breast cancer screening," AJR, Am. J. Roentgenol. **171**, 29–33 (1998).

[2] S. A. Feig, "Increased benefit from shorter screening mammography intervals for women ages 40–49 years," Cancer **80**, 2035–2039 (1997).

[3] L. Tabar, B. Vitak, H. H. Chen, M. F. Yen, S. W. Duffy, and R. A. Smith, "Beyond randomized trials: organized mammographic screening substantially reduces breast cancer mortality," Cancer **91**, 1724–1731 (2001).

[4] S. K. Goergen, J. Evans, G. P. Cohen, and J. H. MacMillan, "Characteristics of breast carcinomas missed by screening radiologists," Radiology **204**, 131–135 (1997).

[5] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," Radiology **219**, 192–202 (2001).

[6] J. E. Meyer and D. B. Kopans, "Stability of a mammographic mass: a false sense of security," AJR, Am. J. Roentgenol. **137**, 595–598 (1981).

[7] E. A. Sickles, "Breast masses: mammographic evaluation," Radiology **173**, 297–303 (1989).

[8] H. K. Hussain, Y. Y. Ng, C. A. Wells, I. B. Nockler, O. M. Curling, R. Carpenter, and N. M. Perry, "The significance of new densities and microcalcification in the second round of breast screening," Clin. Radiol. **54**, 243–247 (1999).

[9] S. Sanjay-Gopal, H. P. Chan, T. Wilson, M. Helvie, N. Petrick, and B. Sahiner, "A regional registration technique for automated interval change analysis of breast lesions on mammograms," Med. Phys. **26**, 2669–2679 (1999).

[10] G. Hermann, R. J. Keller, P. Tartter, L. Bleiweiss, and J. G. Rabinowitz, "Interval changes in nonpalpable breast lesions as an indication of malignancy," Can. Assoc. Radiol. J. **46**, 105–110 (1995).

[11] C. L. Carter, C. Allen, and D. E. Henson, "Relation of lesion size, lymph node status, and survival in 24,740 breast cancer cases," Cancer **63**, 181–187 (1989).

[12] G. D'Eredita', C. Giardina, M. Martellotta, T. Natale, and F. Ferrarese, "Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution," Eur. J. Cancer **37**, 591–596 (2001).

[13] J. A. Baker, P. J. Kornguth, and C. E. Floyd, Jr., "Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description," Am. J. Roentgenol. **166**, 773–778 (1996).

[14] L. Liberman and J. H. Menell, "Breast imaging reporting and data system (BI-RADS)," Radiol. Clin. North Am. **40**, 409–430 (2002).

[15] C. J. Vyborny, M. L. Giger, and R. M. Nishikawa, "Computer-aided detection and diagnosis of breast cancer," Radiol. Clin. North Am. **38**, 725–740 (2000).

[16] I. Leichter, S. Buchbinder, P. Bamberger, B. Novak, S. Fields, and R. Lederman, "Quantitative characterization of mass lesions on digitized mammograms for computer-assisted diagnosis," Invest. Radiol. **35**, 366–372 (2000).

[17] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," Med. Phys. **25**, 516–526 (1998).

[18] L. Li, Y. Zheng, L. Zhang, and R. A. Clark, "False-positive reduction in CAD mass detection using a competitive classification strategy," Med. Phys. **28**, 250–258 (2001).

[19] B. Zheng, Y. H. Chang, and D. Gur, "On the reporting of mass contrast in CAD research," Med. Phys. **23**, 2007–2009 (1996).

[20] W. F. Good, B. Zheng, Y. H. Chang, X. H. Wang, and G. Maitz, "Procrustean image deformation for bilateral subtraction of mammograms," Proc. SPIE **3661**, 1562–1573 (1999).

[21] O. Pawluczyk and M. J. Yaffe, "Field nonuniformity correction for quantitative analysis of digitized mammograms," Med. Phys. **28**, 438–444 (2001).

[22] J. M. Fitzpatrick, "The existence of geometrical density—Image transformations corresponding to object motion," Comput. Vis. Graph. Image Process. **44**, 155–174 (1988).

[23] B. J. McParland, "A comparison of two mammography film-screen combinations designed for standard-cycle processing," Br. J. Radiol. **72**, 73–75 (1999).

[24] T. Matsumoto, H. Yoshimura, K. Doi, M. L. Giger, A. Kano, H. MacMahon, K. Abe, and S. M. Montner, "Image feature analysis of false-positive diagnoses produced by automated detection of lung nodules," Invest. Radiol. **27**, 587–597 (1992).

[25] R. M. Nishikawa, M. L. Giger, K. Doi, C. E. Metz, F. F. Yin, C. J. Vyborny, and R. A. Schmidt, "Effect of case selection on the performance of computer-aided detection schemes," Med. Phys. **21**, 265–269 (1994).

Performance Change of Mammographic CAD Schemes Optimized with Most-Recent and Prior Image Databases[1]

Bin Zheng, PhD, Walter F. Good, PhD, Derek R. Armfield, MD, Cathy Cohen, MD
Todd Hertzberg, MD, Jules H. Sumkin, DO, David Gur, ScD

Rationale and Objectives. The authors evaluated performance changes in the detection of masses on "current" (latest) and "prior" images by computer-aided diagnosis (CAD) schemes that had been optimized with databases of current and prior mammograms.

Materials and Methods. The authors selected 260 pairs of matched consecutive mammograms. Each current image depicted one or two verified masses. All prior images had been interpreted originally as negative or probably benign. A CAD scheme initially detected 261 mass regions and 465 false-positive regions on the current images, and 252 corresponding mass regions (early signs) and 471 false-positive regions on prior images. These regions were divided into two training and two testing databases. The current and prior training databases were used to optimize two CAD schemes with a genetic algorithm. These schemes were evaluated with two independent testing databases.

Results. The scheme optimized with current images produced areas under the receiver operating characteristic curve of 0.89 ± 0.01 and 0.65 ± 0.02 when tested with current images and prior images, respectively. The scheme optimized with prior images produced areas under the receiver operating characteristic curve of 0.81 ± 0.02 and 0.71 ± 0.02 when tested with current images and prior images, respectively. Performance changes for both current and prior testing databases were significant ($P < .01$) for the two schemes.

Conclusion. CAD schemes trained with current images do not perform optimally in detecting masses depicted on prior images. To optimize CAD schemes for early detection, it may be important to include in the training database a large fraction of prior images originally reported as negative and later proven to be positive.

Key Words. Breast neoplasms, diagnosis; breast radiography; computers, diagnostic aid.

© AUR, 2003

Mammography is considered the most reliable and cost-effective screening method for the early detection of breast cancers, which could lead to early treatment and substantially reduce associated mortality and morbidity (1,2). The large volume of mammograms obtained and the low cancer detection rates in a mammographic screening environment could result in radiologists missing as many as 10%–30% of cancers rated "visible" during retrospective reviews (3,4). To assist radiologists in detecting more cancers at screening, computer-aided detection (CAD) systems are being used in many medical institutions around the world (5,6). A number of studies have been conducted to assess their possible effect on radiologists' performance. Although there is no general agreement on whether and how CAD systems help radiologists

improve their diagnostic accuracy (7,8), a number of studies have demonstrated that the performance of the particular CAD scheme (including sensitivity, false-positive rate, and reproducibility) could be important in this regard (9–11).

Current guidelines recommend periodic mammographic screening for women over age 40 years (12). As compliance increases in the general population, a large fraction of patients will have undergone a series of mammographic examinations. As more of the most easily detected cancers are identified during the initial examination with the incorporation of CAD into the diagnostic process, detected breast cancers will be shifted, on average, toward an earlier stage. In other words, more subtle cancers will be considered visible or detectable on routine mammograms. This will occur also in part because of the availability of previous images for comparison, which could help radiologists detect more subtle cancers (13,14). In this changing environment, it is not clear whether current CAD schemes optimized with a large number of easily detected cancers are best suited for the detection of earlier or more subtle cancers. This may become an important issue in developing and evaluating new CAD schemes. In our experiment, an artificial neural network (ANN) previously used in our own CAD scheme for mass detection was reoptimized separately by means of mass regions depicted on "current" images (from the most recent examination, at which the mass was actually reported, leading to biopsy) and those depicted on the corresponding "prior" images (originally interpreted as negative). Hence, two different schemes were used. The changes in their performance were then evaluated when they were applied to independent sets of cases with masses depicted on both current and prior images.

MATERIALS AND METHODS

We searched our database for verified cases in which both current and prior images had been collected and digitized. Inclusion criteria required that at least one mass had been identified by a radiologist on the current images and that biopsy had been performed as a result. In addition, during a retrospective review and with the support of available source documents, an experienced observer (B.Z.) had to be able to identify a mass at the corresponding locations on the prior images. In each case, the most recent prior image had been interpreted as negative or "not highly suspicious."

As a result, 134 cases were selected for this study. The mass was visible on both views (craniocaudal and mediolateral oblique) in 126 cases and on only one view in eight cases. Hence, 260 pairs of images, with each pair consisting of one current image and one prior image, were included in the study. On these images 270 distinct mass regions were identified (10 images depicted two mass regions), 220 of which were associated with biopsy-proved malignancy (50 were benign). The locations of all masses depicted on current images and the corresponding regions on prior images were visually identified as confirmed by the diagnostic reports and pathology results. The centerpoint (x,y coordinate) of each verified mass region was marked manually and saved in a reference (or "truth") file.

All 520 images (260 current and 260 prior) were processed by a CAD scheme developed previously in our laboratory to identify and classify suspicious regions (15). The scheme includes three stages. First, it uses image subtraction and threshold results after processing by two Gaussian filters with a large difference in kernel sizes (7 pixels and 51 pixels) to search for the initial set of suspicious regions, a process that usually results in the identification of 10–30 suspicious regions per image. In the second stage, on the basis of local contrast measurement, the scheme uses an adaptive region growth algorithm to define three topographic layers for each region. Through the imposition of threshold conditions of growth ratio and shape factor for each layer in the regions identified as potential lesions, this stage eliminates approximately 85% of identified regions from consideration, while maintaining high sensitivity. A set of features is computed for each detected region. During the third stage, the remaining regions are classified according to scores generated by a nonlinear multilayer feature–based classifier, defining the likelihood of there being true-positive findings in those regions (16).

In this experiment, all remaining regions identified as suspicious mass regions after the second stage of the CAD scheme were selected for further consideration (the classification scores in the third stage were ignored). As a result, 726 suspicious regions on the 260 current images and 723 suspicious regions on the 260 prior images were selected. If the location of a selected region matched that of a verified mass, the region identification was considered true-positive. Specifically, the distance between the center of gravity of a region, as detected automatically by the CAD scheme, and the center of the mass, as recorded in the reference file, had to be shorter than the radius of

Number of Suspicious Mass Regions in Each Data Set

Images	Training Data Set		Testing Data Set	
	True-Positive	False-Positive	True-Positive	False-Positive
Current	131 (103)	233	130 (108)	232
Prior	126 (100)	236	126 (104)	235

Note.—Numbers in parentheses indicate the regions associated with malignant masses.

the longest axis of the detected region. Otherwise, the region was considered a false-positive identification.

The locations of 261 of the 726 selected regions on current images matched those of verified masses, compared with 252 of the 723 regions on the prior images. All true-positive and false-positive regions were then randomly divided into four mutually exclusive data sets, two for current images and two for prior images. To minimize potential bias, true-positive regions of the same mass (depicted on craniocaudal and mediolateral oblique views) were assigned to the same data set (either training or testing), and when a mass region was assigned to the training (or testing) subset in current images, its corresponding regions as depicted on prior images were also assigned to the training (or testing) subset. The Table summarizes the number and distribution of true-positive regions and false-positive regions in each of the four data sets.

Training data sets from the current and prior images were used to optimize two feature-based ANNs independently as substitutes for the third stage in our CAD scheme (16). Previous studies have demonstrated that the feature distributions were different for mass regions depicted on current images and those depicted on prior images and that different feature sets should be used for optimal classification results (17,18). Therefore, we applied a genetic algorithm to search separately for optimal sets of features on current images and on prior images, using the genetic algorithm software and optimization protocol that had been used in our previous studies to optimize both Bayesian belief networks (19) and ANNs (20).

In brief, a binary coding method is applied to create a chromosome used in the genetic algorithm. Each extracted feature corresponds to a gene (that is, either to 0 or to 1). To determine the optimal number of neurons in the second (hidden) layer of the ANN, we include four additional genes in the chromosome. Hence, the chromosome has a fixed length of 40 genes, of which the first 36 represent extracted image features and the last four indi-

cate the binary-coded number of hidden neurons (eg, 0101 is the code for five hidden neurons) (20). To set up initial parameters in the genetic algorithm software, we included a population size of 100 and assigned the crossover rate, the mutation rate, and the generation gap to 0.6, 0.001, and 1.0, respectively. To minimize overfitting and increase robustness of the ANN performance, we adopted a limited number of training iterations (1,000), as well as a large ratio between the momentum (0.8) and learning rate (0.01) in the ANN. The output of the ROCFIT software program (University of Chicago, Ill) (21) was interfaced with the fitness function of the genetic algorithm, and A_z values computed by the program were defined as fitness criteria in the genetic algorithm. The genetic algorithm was terminated when it either converged to the "highest" A_z value (with no further improvement accomplished in the new generation) or reached a predetermined number of generations (eg, 100).

Using this approach, we generated two optimal ANNs, each using a different training data set. ANN-1 was trained with the suspicious mass regions extracted solely from the current images, and ANN-2 was trained with regions extracted solely from the prior images. Then we applied each of the ANNs to the two mutually exclusive testing data sets of regions extracted from both current and prior images. The classification scores in each test were used to generate four receiver operating characteristic (ROC) curves. The four A_z values were compared. We defined the threshold as a false-positive detection rate similar to that of the leading commercial CAD products—approximately 0.4 false-positive mass regions per image (7). At this level, we found the corresponding detection sensitivity levels and computed the expected number of detected true-positive regions (130 in the data set of current images, and 126 in that of prior images). Thus, we compared the change in expected true-positive detection levels with the use of ANN-1 and ANN-2 for current and prior images at an operating point currently accepted in clinical CAD.

RESULTS

From the genetic algorithm and training data sets of current images and of prior images, two optimal ANNs were generated. ANN-1 included 13 features, and ANN-2 included 11 (Fig 1); four features were common to both. Many of the features are not orthogonal, which is not unique to our scheme. The highest A_z values achieved for the training data sets were 0.92 ± 0.01 for ANN-1 and

Figure 1. Features selected by means of the genetic algorithm for ANN-1 and ANN-2. Those in boldface are common to both ANNs.

ANN-1	ANN-2
1. **Region size (1st layer)**	1. **Region size (1st layer)**
2. Contrast (1st layer)	2. Minimum pixel value inside the region
3. Standard deviation of pixel values (2nd layer)	3. Size growth ratio between 2nd and 3rd layers
4. Circularity (2nd layer)	4. Skewness of pixel values (3rd layer)
5. Region size (3rd layer)	5. Standard deviation of pixel values in background
6. Contrast (3rd layer)	6. Region perimeter divided by size (3rd layer)
7. **Standard deviation of radial length (3rd layer)**	7. **Standard deviation of radial length (3rd layer)**
8. **Circularity (3rd layer)**	8. **Circularity (3rd layer)**
9. Ratio between the maximum and minimum radial lengths (3rd layer)	9. Skewness of pixel values of background
10. Difference of minimum pixel values inside and outside of the growth region (3rd layer)	10. Average local pixel value fluctuation (within a 5 x 5 frame) of the segmented breast area
11. **Region conspicuity (3rd layer)**	11. **Region conspicuity (3rd layer)**
12. Standard deviation of pixel values (3rd layer)	
13. Standard deviation of pixel values in the segmented breast area	

0.76 ± 0.02 for ANN-2. When ANN-1 was applied to the testing data sets, the A_z values were 0.89 ± 0.01 and 0.65 ± 0.02 for current and prior images, respectively. Figure 2 shows three ROC curves for training and two testing results. When ANN-2 was applied to the same data sets, the A_z values were 0.81 ± 0.02 for current and 0.71 ± 0.02 for prior images. Figure 3 shows the corresponding ROC curves for ANN-2.

The test results differed significantly ($P < .01$) between ANN-1 and ANN-2 for both the current and prior image testing data sets. As shown in Figure 4, A_z values were reduced by 9.0% (from 0.89 with ANN-1 to 0.81 with ANN-2) for the current testing data set and increased by 9.2% (from 0.65 to 0.71) for the prior testing data set. In addition, at an operating point of 0.4 false-positive detections per image, the sensitivity levels represented by the two ROC curves in Figure 2 are 0.82 and 0.40. In Figure 3, the corresponding sensitivity levels are 0.68 and 0.52. If we convert these levels to an expected number of detected true-positive mass regions, ANN-1 would detect 18 additional mass regions in the current testing data set, while ANN-2 would detect 15 additional mass regions in the prior testing data set.

The results are not substantially different when benign masses are excluded from the analysis. ANN-1
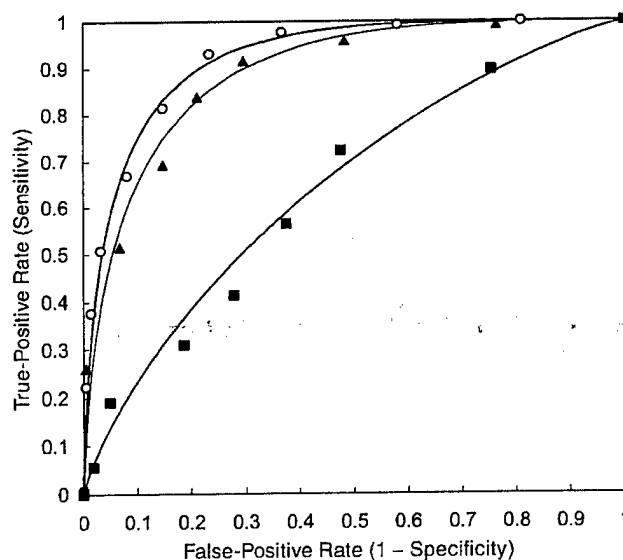

Figure 2. ROC curves showing the performance of ANN-1 during training with the current image data set (O) and during testing with the current image data set (▲) and the prior image data set (■).

yielded performance levels (A_z) of 0.88 ± 0.02 and 0.63 ± 0.02 for current and prior images, respectively; the comparable values for ANN-2 were 0.81 ± 0.02 and 0.70 ± 0.03.
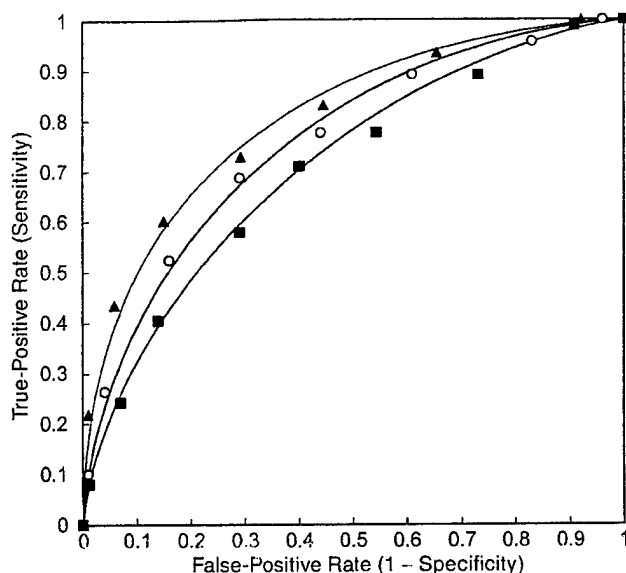
Figure 3. ROC curves showing the performance of ANN-2 during training with the prior image data set (O) and during testing with the current image data set (▲) and the prior image data set (■).

Figure 4. Differences in area under the ROC curve (A_z) for ANN-1 and ANN-2 when tested with the current image data set (▲) and the prior image data set (■).

DISCUSSION

Feature-based machine learning classifiers, such as ANNs, are widely used in CAD schemes as a final stage in identifying and classifying abnormalities. Since these classifiers are trained to generate a "global" function to cover the entire instance space (22), their performance depends heavily on the training databases. This is particularly true in mammography, for which the size and diversity of training data sets are often limited (23,24). Optimal feature sets such as those selected by the genetic algorithm could differ for different limited-size training databases. Hence, the features selected in this study for the current images were very similar but not identical to those selected in our previous studies (16,18). A single CAD scheme that achieves high sensitivity for both subtle and relatively easy-to-detect masses at an acceptable false-positive rate can be developed if a large and diverse image database is available. However, the creation of such a database is very difficult, because image features (including texture- and morphology-based features) are substantially different for suspicious mass regions extracted from current and prior images, as previous studies have demonstrated (17,18).

The CAD scheme trained with the current image data set did not perform optimally when tested with the prior image data set, and vice versa. On the one hand, it is im-
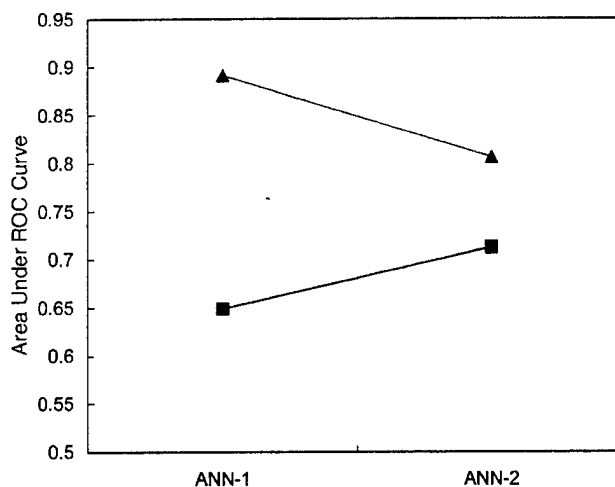
portant for a CAD scheme to detect more subtle masses, because most radiologists can identify the easily detected ones. On the other hand, users may lose confidence in a scheme if it frequently misse⁻ masses that should be easy to detect. Without such c⌐⌐⌐⌐ence, radiologists will most likely be reluctant to accept CAD cuing on subtle masses or make any changes in their initial interpretation (8), preventing the full benefit of CAD schemes from being realized in clinical environments. When ANN-2, which had been trained with the prior image data set, was tested with the current image data set, the testing results were better (higher A_z) than the training results, demonstrating the general robustness of the scheme (Fig 3).

Like most commercially available CAD systems, our CAD scheme was designed to detect, not classify, suspicious abnormalities. Therefore, we believe that the scheme should be highly sensitive to all suspicious mass regions considered "actionable" by radiologists (eg, recommended for follow-up or biopsy), even if some regions later prove benign. One of our previous studies suggested that radiologists' performance in classifying abnormalities as benign or malignant was not affected by the performance of CAD cuing for detection purposes (11). In any event, the inclusion of the benign mass regions as true-positive cases in this experiment did not affect our results and conclusions.

With improvements in diagnostic technology and increasing compliance with screening recommendations among women generally, radiologists have to detect increasingly subtle abnormalities depicted on mammograms.

As a result, the performance of a CAD system that initially provided satisfactory cuing results when optimized could deteriorate substantially over time. Therefore, it may be beneficial to update training data sets periodically and reoptimize the schemes by using a large fraction of new cases originally rated negative and later found positive. An alternative approach could be to provide two types of cues, one trained with current and one with prior images ("early signs"). We believe that our experimental results are not unique to our own image database, our CAD scheme, or ANN-based CAD schemes but should apply to all types of CAD schemes in which feature-based machine learning classifiers are used.

REFERENCES

1. Tabar L, Vitak B, Chen HH, et al. Beyond randomized trials: organized mammographic screening substantially reduces breast cancer mortality. Cancer 2001; 91:1724–1731.
2. Feig S. Increased benefit from shorter screening mammography intervals for women ages 40–49 years. Cancer 1997; 80:2035–2039.
3. Yankaskas BC, Schell MJ, Bird RE, Desrochers DA. Reassessment of breast cancers missed during routine screening mammography: a community-based study. AJR Am J Roentgenol 2001; 177:535–541.
4. Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001; 219:192–202.
5. Malich A, Marx C, Facius M, Böhm T, Fleck M, Kaiser WA. Tumour detection rate of a new commercially available computer-aided detection system. Eur Radiol 2001; 11:2454–2459.
6. Lechner M, Nelson M, Elvecrog E. Comparison of two commercially available computer-aided detection (CAD) systems. Appl Radiol 2002; 31:31–35.
7. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001; 220:781–786.
8. Moberg K, Bjurstam N, Wilczek B, Rostgard L, Egge E, Muren C. Computed assisted detection of interval breast cancers. Eur J Radiol 2001; 39:104–110.
9. te Brake GM, Karssemeijer N, Hendriks JH. Automated detection of breast carcinomas not detected in a screening program. Radiology 1998; 207:465–471.
10. Malich A, Azhari T, Böhm T, Fleck M, Kaiser WA. Reproducibility: an important factor determining the quality of computer aided detection (CAD) systems. Eur J Radiol 2000; 36:170–174.
11. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic reading with different computer-assisted detection cuing environments: preliminary findings. Radiology 2001; 221:633–640.
12. Feig SA, D'Orsi CJ, Hendrick RE. American College of Radiology guidelines for breast cancer screening. AJR Am J Roentgenol 1998; 171:29–33.
13. Bassett LW, Shayestehfar B, Hirbawi I. Obtaining previous mammograms for comparison: usefulness and cost. AJR Am J Roentgenol 1994; 163:1083–1086.
14. Callaway MP, Boggis CR, Astley SA. Influence of previous films on screening mammographic interpretation and detection of breast carcinoma. Clin Radiol 1997; 52:527–529.
15. Zheng B, Chang YH, Gur D. Computerized detection of masses in digitized mammograms using single image segmentation and a multilayer topographic feature analysis. Acad Radiol 1995; 2:959–966.
16. Zheng B, Sumkin JH, Good WF, Maitz GS, Chang YH, Gur D. Applying computer-assisted detection schemes to digitized mammograms after JPEG data compression: an assessment. Acad Radiol 2000; 7:595–602.
17. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan M. Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign breast masses. Med Phys 2001; 28:2309–2317.
18. Zheng B, Shah R, Wallace L, Hakim C, Ganott MA, Gur D. Computer-aided detection in mammography: an assessment of performance on current and prior images. Acad Radiol 2002; 9:1245–1250.
19. Zheng B, Chang YH, Wang XH, Good WF, Gur D. Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm. Acad Radiol 1999; 6:327–332.
20. Zheng B, Chang YH, Good WF, Gur D. Performance gain in computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering. Med Phys 2001; 28:2302–2308.
21. Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med 1998; 17:1033–1053.
22. Towell G, Shavlik J. An approach to combining explanation-based and neural learning algorithms. Connection Sci 1989; 1:233–255.
23. Zheng B, Chang YH, Good WF, Gur D. Adequacy testing of training set sample sizes in the development of a computer-assisted diagnosis scheme. Acad Radiol 1997; 4:497–502.
24. Kupinski MA, Giger ML. Feature selection with limited datasets. Med Phys 1999; 26:2176–2182.

Recall and Detection Rates in Screening Mammography

A Review of Clinical Experience—Implications for Practice Guidelines

David Gur, Sc.D.[1]
Jules H. Sumkin, D.O.[1]
Lara A. Hardesty, M.D.[1]
Ronald J. Clearfield, M.D.[1]
Cathy S. Cohen, M.D.[1]
Marie A. Ganott, M.D.[1]
Christiane M. Hakim, M.D.[1]
Kathleen M. Harris, M.D.[1]
William R. Poller, M.D.[1]
Ratan Shah, M.D.[1]
Luisa P. Wallace, M.D.[1]
Howard E. Rockette, Ph.D[1,2]

[1] Department of Radiology, University of Pittsburgh and Magee-Womens Hospital, Pittsburgh, Pennsylvania.

[2] Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania.

BACKGROUND. The authors investigated the correlation between recall and detection rates in a group of 10 radiologists who had read a high volume of screening mammograms in an academic institution.

METHODS. Practice-related and outcome-related databases of verified cases were used to compute recall rates and tumor detection rates for a group of 10 Mammography Quality Standard Act (MQSA)-certified radiologists who interpreted a total of 98,668 screening mammograms during the years 2000, 2001, and 2002. The relation between recall and detection rates for these individuals was investigated using parametric Pearson (r) and nonparametric Spearman (rho) correlation coefficients. The effect of the volume of mammograms interpreted by individual radiologists was assessed using partial correlations controlling for total reading volumes.

RESULTS. A wide variability of recall rates (range, 7.7–17.2%) and detection rates (range, 2.6–5.4 per 1000 mammograms) was observed in the current study. A statistically significant correlation ($P < 0.05$) between recall and detection rates was observed in this group of 10 experienced radiologists. The results remained significant ($P < 0.05$) after accounting for the volume of mammograms interpreted by each radiologist.

CONCLUSIONS. Optimal performance in screening mammography should be evaluated quantitatively. The general pressure to reduce recall rates through "practice guidelines" to below a fixed level for all radiologists should be assessed carefully. *Cancer* 2004;100:1590–4. © *2004 American Cancer Society.*

KEYWORDS: mammography, screening, tumor detection rates, recall rates.

As periodic mammographic screening is rapidly gaining acceptance, our understanding of many strategic, operational, and financial issues related to this practice is improving as well. Several performance indices have been used to define "optimal" practice parameters in screening mammography. These include, but are not limited to, sensitivity, specificity, positive predictive value (PPV), and cost per detected tumor.[1,2] Clearly, the focus of screening for early detection should primarily be on improved sensitivity. At the same time, the large number of patients being recalled for additional procedures as a result of an initial review is a recognized problem for the very same reasons (operational and financial), with the added concern of the well documented increase of anxiety levels in women who are recalled.[3,4] Therefore, there is a belief that through a variety of actions including but not limited to specific and targeted training, one can augment observer performance levels, including the reduction of recall rates in screening mammography.[5,6] Although not specifically

regulated, there is a publicly stated goal to reduce recall levels to < 10%.[5,7] The question of what effect, if any, does a forced reduction in recall rates have on detection rates remains somewhat controversial. Some studies suggest that recall and detection rates are not highly correlated (particularly at high recall rates); hence, a reduction in the former does not necessarily affect the latter.[6,8] Other researchers believe that, after appropriate training, highly experienced radiologists individually operate largely along a single receiver operating characteristic curve; hence, pressuring them to reduce their recall rate may result in a corresponding reduction in the detection rates as well.[2,9] Because of the well documented variability among radiologists, the latter effect and its possible magnitude have to our knowledge been investigated only recently.[10–13] This type of an investigation is not easy to perform, because the expected yield (detection of actually positive cases that result from the screening) has been reported to be quite low in a population of women who already have been screened in the past.[14,15] Therefore, one generally needs to evaluate detection rates from the data of large groups of individual radiologists pooled together or have access to sufficient data from radiologists who each have interpreted a large number of mammograms. In this article, we present an analysis of the latter type of investigation.

MATERIALS AND METHODS

Screening mammography examinations performed in the study facilities at Magee-Womens Hospital (of the University of Pittsburgh Medical Center) and its five satellite breast imaging clinics during the years 2000, 2001, and 2002 were reviewed under an Institutional Review Board-approved protocol. Mammograms that had been interpreted by the 10 highest volume mammographers at the study institution during this period were included in the current study.

The data sources used in the current analysis were databases of procedure scheduling, procedure completion, radiology reporting, and procedure-related outcomes as determined from pathology reports. These databases have been assembled from original reports for several reasons, including quality assurance purposes that are required by the Mammography Quality Standard Act (MQSA).[16,17] The computerized reporting system and data entry protocols used in our practice remained the same throughout the study period. Because the number of positive findings leading to the detection of tumors by each individual were low, the records of all mammograms read by each of the participating radiologists "with" and "without" the availability of results from a commercial Computer-Assisted Detection (CAD) system were pooled for the purpose of this analysis. Our clinical practice for screening mammography during this period was film based, and most screening mammograms were read at the main facility in a batch mode. We included in the current analysis the results from the interpretations of the 10 highest volume radiologists in our practice, most of whom were with the study institution throughout much of the period in question. Each has performed > 3500 interpretations of screening mammography examinations.

Recall rates for each radiologist were computed directly from mammography interpretation records (Breast Imaging Reporting and Data System Atlas [BI-RADS® Atlas; American College of Radiology, Reston, VA] rating of 0). We excluded recommendations for recall due to technical reasons ("technical recalls"). These account for approximately 1% of cases. However, recalls resulting from palpable findings during clinical breast examinations were included because the majority of these findings also were depicted in the mammograms. These findings amount to < 1% of examinations; therefore, the underlying rates attributable to mammography interpretations alone are accordingly somewhat lower than those reported in the current study. The effect of "palpable" findings on individual radiologists is expected to be distributed proportionally to their overall volume.

In our practice, the interpretation of some examinations (< 4%) is delayed because of missing comparison films during the initial interpretation. These generally are distributed proportionally to the volume read by each radiologist and are included in the recall rates because it is not clear how many of these cases would have been actually recalled in any case.

Tumor detection rates were computed as follows. We identified the latest screening examination for each detected tumor that resulted in a diagnostic follow-up (recall) and ultimately resulted in pathologically verified carcinoma. The radiologist who interpreted the screening mammogram that led to the detection of breast carcinoma was credited with the finding for the purposes of the current analysis. Cases were excluded from the analysis if the latest screening mammogram prior to biopsy had been performed > 180 days earlier. In our experience, these women generally are "lost" to follow-up at other institutions or ignore the recommendations for a diagnostic workup (recall) altogether. Cancer patients who were referred to us from other facilities and for whom the diagnosis did not originate from a screening examination in one of our facilities were excluded. Women who originally were presented as screening procedures but were diagnosed using additional radio-

graphic procedures or other modalities (e.g., ultrasound) during the same visit ("conversion" cases from screening to diagnostic) were accounted for and were included in the current analysis. However, because a substantial number of these may originally have been identified as "potentially abnormal" by a technologist (who personally shows the case to a radiologist) during a quality assurance review of the images, we repeated the analysis after excluding this group of cases. For the purpose of these analyses, we assume that any effect due to the performance level of the radiologists who were performing and interpreting the diagnostic procedures during the follow-up visit are distributed in a manner that does not affect the study conclusions. The radiologists could not select the examinations they interpreted in our practice.

The correlation between recall and detection rates was evaluated using both the parametric Pearson (r) and the nonparametric Spearman (rho) correlation coefficients. We also examined the results after partial correction for the total volume of mammograms interpreted by each radiologist during the period in question.

RESULTS

Recall and detection rates for the 10 radiologists whose data were analyzed in the current study were computed. Each performed > 3500 interpretations (range, 3605–16,128 interpretations) during the period in question. We were unable to publish detailed information for individual radiologists without providing individually traceable data because each staff radiologist is aware of the approximate volume of screening examinations they interpreted and their approximate recall rate. These 10 radiologists interpreted a total of 98,668 cases during this time and detected 368 cases of carcinoma. Twenty-six "conversion" cases were included in the analysis. These cases originally were presented as a screening procedure but the patients underwent "follow-up" procedures (e.g., ultrasound) during the same visit (because of a physician being present on site at the time of the visit). A wide range of recall rates (range, 7.7–17.2%) and detection rates (range,2.6–5.4 per 1000 mammograms) was observed. Despite the low number of radiologists (10), when recall and detection rates were compared using the parametric Pearson (r) correlation coefficient, the correlation between the recall and detection rates was significant (r = 0.76; P = 0.01). Similarly, a significant correlation was observed in the group of radiologists using the nonparametric Spearman correlation coefficient (rho = 0.72; P = 0.02). A linear least square fit between the recall and detection rates for the
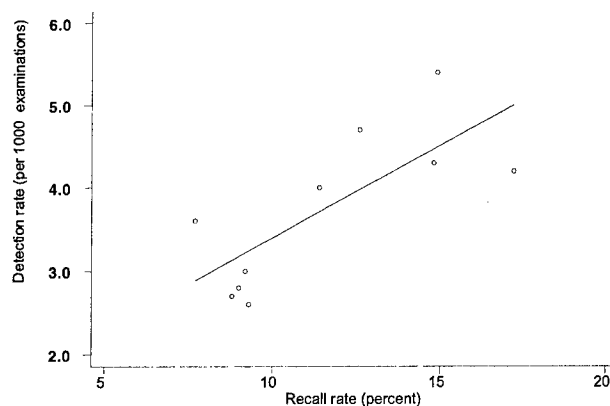

FIGURE 1. A linear fit of detection rates as a function of recall rates for the 10 radiologists in the current study.

group in which each radiologist represents a single "operating point" is presented in Figure 1. Despite significant interreader variability, the slope indicates an average of 0.22 additional detections per 1% increase in recall rates (95% confidence interval on the slope is +0.068 to +0.378). The correlation between recall and detection rates remained significant (P < 0.05) after accounting for the total volume read by each radiologist using partial correlations. Repeated analyses after the exclusion of the 26 "conversion" cases indicated no substantial difference in the correlations reported herein. The correlations remained significant when the analysis was repeated for the 7 (P = 0.05), 8 (P < 0.05), and 9 (P < 0.05) highest volume radiologists. These results demonstrate that, in general, in our practice, the higher the recall rates, the higher the detection rates. This increase in detection rate was found to persist over the range of observed recall rates and extended beyond the currently recommended practice guideline of 10%.

DISCUSSION

There is little doubt that continuing education and training are important factors in the ability of radiologists to be consistent in interpreting mammograms and to improve their overall performance. However, to our knowledge, there are no conclusive data published to date regarding to what extent improvement continues beyond a certain level of training or experience.[12] Although there are questions with regard to whether volume and experience affect performance,[12] the general belief has been that one can reduce recall rates relatively easily without a significant impact on detection rates. As a result, there is an ongoing significant effort to do so, particularly in practices similar to ours with recall

rates that are in the higher range ($\geq$ 10%). PPV as a result of screening has been of great interest as one of the indicators of the performance level of radiologists in this area.[8] However, if sensitivity is affected by recall rates, particularly in a group of well trained, high-volume radiologists whose recall rates are relatively high, the fundamental question of whether to continually pressure them to reduce their recall rates following currently accepted practice guidelines remains. This stems from the fact that the detection of "earlier tumors" with higher recall rates may be as or perhaps more important than actually reducing the recall rates or improving the PPV somewhat. It is interesting to note that an important review of several related issues suggested observations that were similar to those of the current study.[10] Unfortunately, to our knowledge the radiology community has not objectively addressed this potentially important matter to date.

Similar to the findings reported by Yankaskas et al.[8], the results of the current study suggest that detection rates generally are affected by recall rates in the lower range. However, unlike the observations of Yankaskas et al.,[8] the effect in our group of 10 highly trained radiologists, who individually read a reasonably high volume of mammograms, persisted over the entire range of observed recall rates (as high as 17%). In the higher range of recall rates ($\geq$ 7%), Yankaskas et al.[8] showed no correlation between the recall and detection rates. Therefore, their results could suggest that, in this critical range, a reduction in recall rates should not affect the detection rates. It is possible that this difference arises from the fact that the current study took place in a "reasonably stable" screening population in whom the majority of "prevalence (or "baseline") carcinomas" had been detected already. Another possible explanation may be the number of mammograms interpreted by individual radiologists in the two studies. Clearly, more data are needed in this regard.

The total number of mammography screening interpretations by the radiologist with the lowest screening volume reported herein over a 3-year period was relatively low. However, our regionwide referral base was found to result in a large number of other diagnostic and interventional breast-imaging procedures that typically amount to approximately 50% of the screening examinations. Hence, our radiologists should be considered as "specialists" in breast imaging.

It should be noted that in our practice the average recall rates ($\approx$11 percent) are generally relatively high compared with some reports,[18,19] and they are in better agreement with, and in some cases lower than,

others.[15,20,21] We have no simple explanation for this observation. The results of the current study are in agreement with the findings of Beam et al.[12] and others in that there is a large variability in the performance of the radiologists in this area. We did not detect a significant correlation between the volume read by the individual radiologists during the period in question and their performance level, although the radiologists in the current study all can be considered high volume, "well trained" readers with significant experience. There are several arguments one can raise with regard to why the estimated recall and detection rates in the current study may not be precise in terms of absolute values. These include but are not limited to the inclusion of palpable cases and incomplete follow-up of cancer patients who may be lost to other institutions. The fact that our primary area of interest is the relative performance levels of the radiologists (rather than absolute) makes the results valid despite these limitations, as long as one does not bias the interpretation process by selectively assigning a specific subset to be interpreted by one radiologist or another (e.g., all "high risk" women or all examinations of women with dense breasts are assigned to "conservative" or "high-volume" radiologists). This was clearly not the case in our practice. Therefore, one would expect that any related corrections as a result of these limitations would be largely proportional to the volume of cases interpreted by each radiologist in the course of their routine clinical practice. The correlation between detection rates and outcome or even "average stage of disease" at the time of detection is beyond the scope of this project because the number of tumors detected by an individual radiologist was too small and the follow-up time after detection too short to meaningfully assess differences, if any, in outcome.

The results of the current study suggest that before we unilaterally pressure radiologists to reduce their recall rates because of a notion that this will improve our practices (and reduce overall management costs), we need to carefully evaluate the impact such an effort may have on early (and perhaps even "earlier") detection. If we believe that screening should focus primarily on maximizing early detection, and the earlier the better, one has to consider whether there may be an individualized optimal operating level that should be considered, rather than a "globally" recommended practice guideline of a maximum "acceptable" recall rate that applies to all screening mammographers. This view may be supported by women who appear to strongly prefer a small increase in detection rates, even at the expense of higher recall rates and the associated impact in terms of cost and added

anxiety.[22–24] The current limited study included a group of 10 academic radiologists practicing at 1 institution under 1 set of practice conditions. Clearly, more data are required before one can generalize the findings reported herein to the population of radiologists who interpret screening mammography in this country. At the same time, the number and type of examinations used in the current analysis may be generalizable to the screening population in a large number of academic practices around the U.S.

Conclusions

The performance level of a radiologist in the screening environment is a complex, multifactorial issue that cannot and should not be simplified. Reducing recall rates by "decree" (through the enforcement of recommended practice guidelines) may result in a corresponding reduction in the detection rates, hence the associated delays. The impact of external pressure on individual radiologists to reduce their recall rates should be evaluated carefully.

REFERENCES

1. Linver MN. Audits measure practice quality of mammography. *Diagn Imaging.* 2000;22:57–61.
2. Burnside E, Belkora J, Esserman L. The impact of alternative practices on the cost and quality of mammographic screening in the United States. *Clin Breast Cancer.* 2001; 2:145–152.
3. Brett J, Austoker J. Women who are recalled for further investigation for breast screening: psychological consequences 3 years after recall and factors affecting re-attendance. *J Public Health Med.* 2001;23:292–300.
4. Sandin B, Chorot P, Valiente RM, Lostao L, Santed MA. Adverse psychological effects in women attending a second-stage breast cancer screening. *J Psychosom Res.* 2002;52:303–309.
5. Feig SA. Economic challenges in breast imaging. A survivor's guide to success. *Radiol Clin North Am.* 2000;38:843–852.
6. Sickles EA. Successful methods to reduce false-positive mammography interpretations. *Radiol Clin North Am.* 2000; 38:693–700.
7. U.S. Department of Health and Human Services, Agency for Health Care Policy and Research. Clinical practice guideline number 13: quality determinants of mammography. AHCPR Pub. No. 95-0632. Washington, DC: . U.S Department of Health and Human Services, Agency for Health Care Policy and Research, 1994:78–86.
8. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. *AJR Am J Roentgenol.* 2001;177: 543–549.
9. Kopans DB. The accuracy of mammographic interpretation [editorial]. *N Engl J Med.* 1994;331:1521–1522.
10. Moskowitz M. Retrospective reviews of breast cancer screening: what do we really learn from them. *Radiology.* 1996;199:615–620.
11. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med.* 1994;331:1493–1499.
12. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *J Natl Cancer Inst.* 2003;95: 282–290.
13. Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst.* 2002;94:1373–1380.
14. Frankel SD, Sickles EA, Curpen BN, Sollitto RA, Ominsky SH, Galvin HB. Initial versus subsequent screening mammography: comparison of findings and their prognostic significance. *AJR Am J Roentgenol.* 1995;164:1107–1109.
15. Young WW, Destounis SV, Bonaccio E, Zuley ML. Computer-aided detection in screening mammography: Can it replace the second reader in an independent double read? Preliminary results of a prospective double blinded study. Presented at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America. *Radiology.* 2002;225(P):600.
16. Food and Drug Administration. Quality Standards and Certification Requirements for Mammography Facilities (21 CFR Part 900) *Federal Register* 58, no. (December 21, 1993): 67565.
17. Linver MN, Osuch JR, Brenner RJ, et al. The mammography audit: a primer for the mammography quality standards act (MQSA). *AJR Am J Roentgenol.* 1995;165:19–25.
18. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology.* 2002;224:861–869.
19. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology.* 2001; 220:781–786.
20. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. *AJR Am J Roentgenol.* 2003;180:1461–1467.
21. Fletcher SW, Elmore JG. Clinical practice. Mammographic screening for breast cancer. *N Engl J Med.* 2003;348:1672–1680.
22. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch H. US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. *BMJ.* 2000;320:1635–1640.
23. Nekhlyudov L, Ross-Degnan D, Fletcher SW. Beliefs and expectations of women under 50 years old regarding screening mammography: a qualitative study. *J Gen Intern Med.* 2003;18:182–189.
24. Silverman E, Woloshin S, Schwartz LM, Byram SJ, Welch HG, Fischhoff B. Women's views on breast cancer risk and screening mammography: a qualitative interview study. *Med Decis Making.* 2001;21:231–240.

ARTICLES

Changes in Breast Cancer Detection and Mammography Recall Rates After the Introduction of a Computer-Aided Detection System

David Gur, Jules H. Sumkin, Howard E. Rockette, Marie Ganott, Christiane Hakim, Lara Hardesty, William R. Poller, Ratan Shah, Luisa Wallace

Background: Computer-aided mammography is rapidly gaining clinical acceptance, but few data demonstrate its actual benefit in the clinical environment. We assessed changes in mammography recall and cancer detection rates after the introduction of a computer-aided detection system into a clinical radiology practice in an academic setting. *Methods:* We used verified practice- and outcome-related databases to compute recall rates and cancer detection rates for 24 Mammography Quality Standards Act–certified academic radiologists in our practice who interpreted 115 571 screening mammograms with (n = 59 139) or without (n = 56 432) the use of a computer-aided detection system. All statistical tests were two-sided. *Results:* For the entire group of 24 radiologists, recall rates were similar for mammograms interpreted without and with computer-aided detection (11.39% versus 11.40%; percent difference = 0.09, 95% confidence interval [CI] = −11 to 11; P = .96) as were the breast cancer detection rates for mammograms interpreted without and with computer-aided detection (3.49% versus 3.55% per 1000 screening examinations; percent difference = 1.7, 95% CI = −11 to 19; P = .68). For the seven high-volume radiologists (i.e., those who interpreted more than 8000 screening mammograms each over a 3-year period), the recall rates were similar for mammograms interpreted without and with computer-aided detection (11.62% versus 11.05%; percent difference = −4.9, 95% CI = −21 to 4; P = .16), as were the breast cancer detection rates for mammograms interpreted without and with computer-aided detection (3.61% versus 3.49% per 1000 screening examinations; percent difference = −3.2, 95% CI = −15 to 9; P = .54). *Conclusion:* The introduction of computer-aided detection into this practice was not associated with statistically significant changes in recall and breast cancer detection rates, both for the entire group of radiologists and for the subset of radiologists who interpreted high volumes of mammograms. [J Natl Cancer Inst 2004;96:185–90]

A mounting body of evidence suggests that early detection of breast cancer through periodic mammography screening reduces the morbidity and mortality associated with this disease *(1,2)*. Mammography screening is rapidly gaining acceptance worldwide, and the number of mammography procedures performed continues to increase *(3,4)*. However, mammography screening has a relatively low cancer detection rate of only two to six cancers per 1000 mammograms after the first 2 years of screening *(5)*.

The performance levels among radiologists who read and interpret mammograms vary widely. Several factors may account for this variability. These include, but are not limited to, the low incidence of breast cancer, the difficulty in identifying suspicious (i.e., potentially malignant) regions in the surrounding breast tissue, and the tedious and somewhat repetitious nature of the task of reading mammograms *(5–7)*.

In recent years, a major effort has been expended to develop computer-aided detection systems to assist radiologists with the diagnostic process. The hope is that these computer-aided detection systems will improve the sensitivity of mammography without substantially increasing mammography recall rates, in addition to possibly decreasing inter-reader variability. These systems are intended for the early detection of breast cancer and, accordingly, are designed to assist the radiologist in the identification (i.e., detection) of suspicious regions (i.e., findings), such as clustered microcalcifications and masses *(8–10)*. Computer-aided diagnosis (discrimination) systems are currently being developed to help radiologists determine whether an identified suspicious region is likely to represent a benign or a malignant finding *(11–13)*.

The U.S. Food and Drug Administration (FDA) has approved several computer-aided detection systems for clinical use, and Medicare and many insurance companies have approved reimbursement for the use of these systems in clinical practice. The initial FDA approval process for these systems included retrospective interpretations of select groups of cases in a laboratory environment *(9,14,15)*. Results of these studies *(9,15)* suggest that the use of computer-aided detection systems can potentially increase cancer detection rates by approximately 20% without substantially increasing recall rates. However, there are only limited data on the impact of such systems when used prospectively in a clinical environment *(16–19)*. We used large, prospectively ascertained databases to evaluate the recall and cancer detection rates in our clinical breast imaging practice in an

academic setting for a 3-year period during which a computer-aided diagnosis system was introduced.

METHODS

Subjects and General Procedures

All screening mammography examinations performed in our facilities at Magee-Womens Hospital of the University of Pittsburgh Medical Center (Pittsburgh, PA) and its five satellite breast imaging clinics during 2000, 2001, and 2002 were included in this study. Our study was carried out under an institutional review board–approved protocol.

The data sources for our analysis were databases that contained information on procedure scheduling, procedure completion, radiology reporting, and procedure-related outcomes as determined from relevant pathology reports. These databases were assembled from the original reports for quality assurance purposes, as required by the Mammography Quality Standards Act (MQSA) (20), among other reasons. The same computerized reporting system was in use throughout the study period.

In the second quarter of 2001, we introduced a computer-aided detection system (R2 Technologies, Los Altos, CA) into our clinical practice at the main facility, where most of the screening mammograms in our practice were read in batch mode. By the third quarter of 2001, more than 70% of the screening mammograms were interpreted with use of the computer-aided detection system. By the fourth quarter of 2001, more than 80% of the screening mammograms were interpreted with the assistance of the computer-aided detection system. The radiologists in our practice could not select which mammograms would be interpreted with or without the computer-aided detection system. After training on the computer-aided detection system was completed (June 2001), all screening mammograms interpreted in our main facility were processed by and interpreted with the assistance of the computer-aided detection system. Radiologists at the five satellite clinics sometimes reviewed screening mammograms if time allowed, but the number of these cases was small, and there was no selection process that could bias the analyses performed in this study. Knowing the schedule for radiologists' presence at the remote sites, we assembled a batch of serially acquired mammograms for them to read in the same way they would be read at the central facility, and those mammograms were interpreted and reported in the same manner (with the exception of the use of computer-aided detection). This set of mammograms was not specifically selected because of suspicious findings by the technologists. To reduce possible biases, an individual not involved in this investigation was asked to examine summaries of time-dependent recall rates for all radiologists in our practice for the study period. A different team examined all cancers detected throughout our practice as a result of screening mammography during the same period.

During the study period, our practice performed a total of 115 571 screening examinations that were interpreted by 24 radiologists, 18 of whom interpreted more than 1000 mammograms each. All radiologists were members of the Breast Imaging Section of the Department of Radiology and would be considered breast imaging specialists in an academic practice. We also repeated our analysis by using only data for the seven highest volume radiologists, all of whom read more than 8000 mammograms each over a 3-year period. These seven radiolo-

gists, who were with our institution throughout the study period, performed the most readings, both with and without computer-aided detection assistance.

For the purpose of computing recall rates, mammograms were considered to be positive if recall for additional imaging evaluation was recommended (i.e., mammograms classified as Breast Imaging Reporting and Data System [BI-RADS] category 0) and negative if a 1-year follow-up was recommended (i.e., mammograms classified as either BI-RADS category 1 or 2) (21). Radiologists at these facilities did not use BI-RADS assessment categories 3, 4, or 5 for screening examinations. Positive outcome was defined as breast cancer detected as a result of the diagnostic work-up initiated by a positive screening mammogram.

Computation of Mammography Recall Rates

Recall rates for each radiologist and for the group of 24 radiologists were computed directly from mammographic interpretation records. In all of our analyses, we excluded recommendations for recall that were due to technical reasons, such as image artifacts (<1%). Recalls due to palpable findings identified during clinical breast examinations performed on all women by the technologist were included in our analyses because the majority of these findings were also marked on the mammograms. Such recalls amounted to approximately 1% of the screening examinations; hence, the underlying rates attributable to mammography interpretations alone are approximately 1% lower than those reported here. The women in this group of recalls are not the same as the group of women with palpable findings discovered by the woman herself or by a physician during a breast physical examination. Women in the latter group were scheduled for diagnostic examinations and were not included in our study. In our practice, palpable findings that are discovered by the technologists are noted during the physical examination and the procedure continues as a screening examination (including the use of computer-aided detection). The interpreting radiologists are aware of the technologists' findings and recall the women for additional procedures as needed. We recognize that this practice may not be a common one. We assumed that the effects of recalling this group of women due to palpable findings, if any, on the recall rates of individual radiologists would be proportional to the overall volume of mammograms read by each radiologist; hence, it should not substantially affect the results.

A small percentage (<4%) of the examinations in our practice classified as BI-RADS category 0 were scheduled for an interpretation at a later date because the needed comparison films were missing during the originally scheduled interpretation. Those cases were distributed proportionally to the volume of mammograms read by each radiologist and were included in the recall rates because it was not clear how many of them would have been recalled anyway.

Each mammography examination was identified in our database as to whether computer-aided detection was used during the interpretation. We therefore analyzed the data according to whether cases were interpreted with computer-aided detection.

Computation of Breast Cancer Detection Rates

Breast cancer detection rates were computed as follows: For every breast cancer detected, we found the most recent screening

mammogram that identified a finding that led to a diagnostic follow-up and ultimately resulted in a biopsy that was positive for cancer. Only the interpreter of the original screening mammogram that led to the detection of breast cancer was credited with the finding (i.e., invasive and ductal carcinoma *in situ*). Findings of lobular carcinoma *in situ* were not attributed to the interpreting radiologist as a cancer detected in the analyses. If a woman was recommended for a biopsy directly as a result of the screening examination, the interpreter was credited with the finding as well. Cases were excluded from the analysis if the most recent screening mammogram prior to biopsy had been performed more than 180 days before the biopsy or if the original interpreter had not recommended a recall (i.e., false-negative cases). We chose a cutoff of 180 days because we have found that, in the vast majority of cases, women are lost to follow-up or ignore the recall recommendation altogether if the recommended follow-up diagnostic procedure is not scheduled within 90 days or performed within 180 days of the original mammogram. We attributed any subsequent findings associated with recalls for diagnostic work-ups that did not take place within 180 days of the original mammogram to the subsequent examination. We included all examinations that that had been originally scheduled as screening procedures but were diagnosed during the same visit and during which a diagnosis was made that resulted in a positive outcome (i.e., converted into a diagnostic procedure that led to a finding of cancer). However, these cancer cases (n = 30) were excluded from the computed breast cancer detection rates in our analysis (both nominator and denominator) because they were all diagnosed by a radiologist without the use of computer-aided detection, and we therefore could not determine whether these cases would have been detected had they undergone routine interpretation (with or without computer-aided detection) as a routine screening procedure. In addition, all breast cancer patients who were referred to us from other facilities and for whom the diagnosis did not originate from a screening examination done at one of our facilities were excluded from the analysis.

Statistical Methods

Recall and detection rates with and without computer-aided detection were compared by using a generalized estimating equations (GEE) logistic regression model that accounts for clustering of findings within each reader *(22)*. In addition, we asked an independent team of investigators to evaluate the numbers of cancer cases that were detected with and without computer-aided detection by the type of abnormality(s) noted in the original report. Those findings were assigned to one of the following categories: 1) mass(es) only; 2) clustered microcalcifications only; 3) mass(es) and clustered microcalcifications; and

4) other findings. Because the performance levels of computer-aided detection systems are generally outstanding for detecting microcalcifications *(16)*, we used the GEE model to analyze our findings with respect to possible changes in the percentage of cancer detections attributable to microcalcification clusters associated with the use of computer-aided detection. In addition, all analyses were repeated using a mixed-effect logistic regression model in which readers were considered a random effect, and modality (i.e., with or without computer-aided detection) was considered a fixed effect *(23)*. We also examined data from the seven high-volume radiologists (i.e., those who interpreted more than 8000 mammograms each during the study period). Because of the serial nature of the analysis (namely, this was not a randomized study), we repeated the analyses with respect to the timing of the major use of computer-aided detection in our practice by comparing the results for all cases interpreted without computer-aided detection from January 1, 2000, through June 30, 2001, when computer-aided diagnosis was used in only a small percentage of cases (<0.2%) at our facilities, with results for all cases interpreted with computer-aided detection from October 1, 2001, through December 31, 2002, when most (>93%) of the cases at our facilities were interpreted with computer-aided detection. All statistical tests were two-sided.

RESULTS

The mean age of the screened population (n = 115 571) during the study period was 50.05 years (standard deviation = 11.17 years). During the study period, the percentage of women who were screened for the first time gradually decreased from approximately 40% in 2000 to 30% in the last quarter of 2002, whereas the percentage of women who had repeated screenings gradually increased.

Table 1 summarizes our data for the 24 radiologists who interpreted screening mammograms at our facility with and without the use of a computer-aided detection system. Among the 115 571 examinations in our database, 56 432 (48.8%) were interpreted without the use of the computer-aided detection system and 59 139 (51.2%) were interpreted with the use of the computer-aided detection system. Recall rates for the entire group of 24 radiologists were 11.39% for mammograms interpreted without computer-aided detection and 11.40% for mammograms interpreted with it (percent difference = 0.09, 95% confidence interval [CI] = −11 to 11; P = .96). Recall rates for the 18 radiologists who interpreted more than 1000 mammograms each during the study period ranged from 7.7% to 17.2% (data not shown). Recall rates for the seven high-volume radiologists who interpreted more than 8000 mammograms each during the study period ranged from 7.7% to 14.9% (data not shown). Among this latter group of radiologists, there was no

Table 1. Mammography recall rates and breast cancer detection rates for 24 radiologists performing screening mammograms without and with computer-aided detection*

Type of interpretation	No. of mammograms read	No. of recalls	No. of breast cancers detected	Recall rate, %	Breast cancer detection rate per 1000 mammograms read
Without computer-aided detection	56 432	6430	197	11.39	3.49
With computer-aided detection	59 139	6741	210	11.40	3.55
Total	115 571	13 171	407	11.40	3.52

*The analysis excluded 30 conversion (screening to diagnostic) cancer cases, all of which were interpreted without computer-aided detection.

statistically significant correlation (rho = −0.21, P = .64) between recall rate and the total number of screening mammograms interpreted by individual radiologists. In our practice, approximately 3.0% of the cases recommended for recall are typically lost to follow-up because the woman either undergoes re-screening at another institution or ignores our recommendations. This group remained relatively constant as a percentage of recalled women over the period in question.

Table 2 summarizes our data for the seven high-volume radiologists who interpreted more than 8000 screening mammograms each with and without the use of a computer-aided detection system. During the study period, these radiologists interpreted a total of 82 129 screening mammograms and were credited with the detection of 292 breast cancers as a result of these screening procedures. In this group, the recall rates decreased from 11.62% for mammograms interpreted without computer-aided detection to 11.05% for mammograms interpreted with computer-aided detection (percent difference = −4.9, 95% CI = −21 to 4; P = .16).

Breast cancer detection rates for the entire group of 24 radiologists were 3.49 per 1000 screening examinations for mammograms interpreted without computer-aided detection and 3.55 per 1000 screening examinations for mammograms interpreted with it (percent difference = 1.7, 95% CI = −11 to 19; P = .68) (Table 1). Breast cancer detection rates for the seven high-volume radiologists were 3.61 per 1000 screening examinations for mammograms interpreted without computer-aided detection and 3.49 per 1000 screening examinations for mammograms interpreted with computer-aided detection (percent difference = −3.2, 95% CI = −15 to 9; P = .54) (Table 2).

The cancer detection rates associated with recalls due to the detection of clustered microcalcifications alone were 1.35 per 1000 mammograms interpreted without computer-aided detection and 1.44 per 1000 mammograms interpreted with computer-aided detection (P = .66) (data not shown). We observed no trend in breast cancer detection rates over time when we reviewed average detection rates for all 24 radiologists by calendar quarter (data not shown). We repeated our analyses using a random-effects logistic regression model and found that there were no statistically significant changes in recall rates or detection rates for all measurements presented above. Our results were not substantially affected when we compared only mammograms interpreted without computer-aided detection prior to July 1, 2001, with only those interpreted with computer-aided detection after October 1, 2001.

DISCUSSION

The introduction of computer-aided detection into our practice was not associated with statistically significant changes in recall and breast cancer detection rates for the entire group of radiologists as well as for the subset of seven radiologists who interpreted high volumes of mammograms. The magnitudes of the improvements we observed were substantially less than those reported in the literature as the range of possible improvements based on retrospective analyses and limited prospective data (9,17,18). The improvements we observed may be attributable to the better detection of clustered microcalcifications associated with malignancy. Our findings are consistent with the range of improvement in detection rates estimated and reported by others (9,16–18). However, our large confidence intervals reflect the relatively low number of breast cancers detected with and without computer-aided detection and the large inter-reader variability among the radiologists in our practice. Because there were no repeat measures in this database—that is, each of the examinations was interpreted only once by one radiologist—we could not assess intra-reader variability.

It should be noted that we could not provide detailed information for individual radiologists without providing individually traceable data because each staff radiologist knows his or her reading volume and approximate recall rate. Our data are not adjusted for any learning effect: namely, the majority of interpretations made without computer-aided detection occurred chronologically prior to those made with computer-aided detection. We also did not account for any effect that may have resulted from a continuous effort to improve performance (in particular, sensitivity) by group reviews of all false-negative cases or from the steps undertaken to reduce recall rates through various actions, such as monthly performance reviews and direct consultation with interpreters who had higher-than-average recall rates.

Although one could argue that some or all of the reduction in recall rates we observed for the high-volume radiologists may be attributable to the use of computer-aided detection, the corresponding decrease in cancer detection rates we observed among the radiologists in this group is not easily explained by expected practice variations. An assessment of whether the small improvement we observed in cancer detection is due to learning effects—namely, that our radiologists had substantially more overall experience interpreting mammograms without computer-aided detection than with computer-aided detection—is beyond the scope of this investigation.

This investigation covered a period during which conventional film mammography was performed in all of our screening procedures. Hence, we cannot comment on the possible effect of computer-aided detection in a digital mammography environment. In our study, we did not account for women who had decided to follow up on our recommendations elsewhere. However, because compliance in patient follow-up was relatively constant during the study period, any bias in the results due to changes in patient loss to follow-up is likely to be small.

There are limited reported data concerning the actual effect of computer-aided detection on breast cancer detection and mam-

Table 2. Mammography recall rates and breast cancer detection rates for the seven high-volume radiologists performing screening mammograms without and with computer-aided detection

Type of interpretation	No. of mammograms read	No. of recalls	No. of breast cancers detected	Recall rate, %	Breast cancer detection rate per 1000 mammograms read
Without computer-aided detection	44 629	5188	161	11.62	3.61
With computer-aided detection	37 500	4145	131	11.05	3.49
Total	82 129	9333	292	11.36	3.56

mography recall rates. The prospective data reported by Freer and Ulissey *(16)*, which suggested a substantial improvement (19.5%) in breast cancer detection rates associated with the use of computer-aided detection systems, may have been affected by the fact that the results of mammographic interpretations without and with computer-aided detection were reported on the same cases (i.e., mammograms were read in one sitting, first without computer-aided detection then immediately afterward with the use of a computer-aided detection system). Another prospective study performed in a similar manner reported a 12% improvement in detection rates associated with the use of a computer-aided detection system *(18)*. This type of protocol, namely reading mammograms without computer-aided detection followed immediately by readings of the same mammograms with the use of a computer-aided detection system and a reassessment of the original finding without computer-aided detection, may have introduced a lower level of vigilance among radiologists during the initial interpretation without computer-aided detection, because they knew that computer-aided detection would be available to them for the final recommendation and that the initial interpretation did not constitute a formal clinical recommendation.

Results of the only study similar to ours, albeit on a substantially smaller group of patients and under a different set of circumstances, suggested that computer-aided detection was associated with a 13% improvement in breast cancer detection rates *(17)*. One of the advantages of the approach taken in our investigation is that the radiologists' interpretations were performed and recorded prospectively in a clinical setting and data were collected primarily for quality-assurance purposes *(24)*.

Our results for the interpretations made with computer-aided detection may be marginally biased because the outcomes of as many as nine recommendations for recalls and three recommendations for biopsies during the last quarter of 2002 are not yet available. Although some of these follow-up procedures or biopsies may ultimately be performed at our institution, we assume that the women who underwent the original mammograms have been lost to follow-up. However, on the basis of our typical recall-to-cancer-detection ratios (approximately 1 of 32 cases) and biopsy-to-confirmed cancer ratios (approximately 1 of 5 cases), we suspect that this bias would not substantially affect our findings or conclusions. It is possible that the gradually increasing fraction of women who had prior screening examinations created a bias in our results. Repeat screening examinations have a slightly lower number of cancers present as more are detected during the first screen, and on average, cancers detected on repeat mammograms may be more "difficult" to detect because more of the "easier" (e.g., larger) cancers are detected during the initial screen. Repeat mammograms have a lower recall rate, as the radiologists have prior films for comparison, to help inform their decision. The availability of prior examinations for comparison (in the repeat examinations) should have aided in the interpretation of these mammograms and offset the possible effect (if any) on the interpretations due to an increase in the "average case difficulty." The fact that our recall rates and detection rates remained virtually constant over time suggests that the possible bias due to a gradual increase in repeat examinations is not a statistically significant factor. We suspect that this increasing availability of prior examinations for comparison is a general phenomenon that is observed by most mammography screening practices and that there is not a simple

way to account for it in an analysis such as the one we performed. When we included the 30 examinations that had been originally scheduled as screening procedures but were diagnosed during the same visit and resulted in a positive outcome in the estimation, our actual cancer detection rate attributable to screening was 3.8 per 1000 examinations, which is reasonable for a population in which the majority of women had undergone several screening procedures prior to the study period *(19)*.

On the basis of published performance levels of other computer-aided detection systems *(25)*, we believe that our results are not unique to the specific computer-aided detection system that is used at our institution. It is possible, however, that in clinical practices with substantially lower recall rates than ours, computer-aided detection would have larger effects on mammography recall rates and detection rates than what we observed. Such an improvement in detection rates would be consistent with results of a study *(17)* that reported lower recall rates without computer-aided detection (8.02%) than with computer-aided detection (8.43%).

The financial implications of our findings are beyond the scope of this work. However, a simple assessment of the additional estimated cost of using computer-aided detection per additional cancer detected in our practice (approximately $150 000 per additional detected cancer, assuming a reimbursement rate of $10 per case for professional and technical components combined) clearly indicates that more rigorous evaluations of the cost effectiveness of this practice are needed.

Our observations with respect to recall and detection rates may be exceptions (stemming from large inter-practice variations) that highlight the need for additional recall and detection rate data from multiple clinical practices and different reading environments. However, until such data clearly demonstrate that our experience is indeed an exception, these results represent an important first step.

This analysis of our practice was designed to assess the changes, if any, that occurred in recall and breast cancer detection rates with the introduction of computer-aided detection. Our results suggest that, in our practice, neither recall rates nor breast cancer detection rates changed with the introduction of this technology at its current level of performance, particularly as related to the detection of abnormalities other than clustered microcalcifications. Due to large confidence intervals, our results are statistically consistent with the possibility of large improvements in cancer detection rates with computer-aided detection. Yet, actually observed changes in our practice were substantially lower than expected. This is not to say that the use of computer-aided detection would not be beneficial or cost-effective in other practices. Rather, we suggest that, at its current level of performance, computer-aided detection may not improve mammography recall or breast cancer detection rates (especially as related to the detection of masses) in academic practices similar to ours that employ specialists for interpreting screening mammograms.

REFERENCES

(1) Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a survey of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2002;137:347–60.

(2) Tabár T, Duffy SW, Vitak B, Chen HH, Prevost TC. The natural history of breast carcinoma: what have we learned from screening? Cancer 1999;86: 449–62.

(3) Hendrick RE, Klabunde C, Grivegnee A, Pou G, Ballard-Barbash R. Technical quality control practices in mammography screening programs in 22 countries. Int J Qual Health Care 2002;14:219–26.

(4) Ng EH, Ng FC, Tan PH, Low SC, Chiang G, Tan KP, et al. Results of intermediate measures from a population-based, randomized trial of mammographic screening prevalence and detection of breast carcinoma among Asian women: the Singapore Breast Screening Project. Cancer 1998;82: 1521–8.

(5) Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224:861–9.

(6) Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331: 1493–9.

(7) Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst 2003;95:282–90.

(8) Doi K, Giger ML, Nishikawa RM, Schmidt RA. Computer-aided diagnosis of breast cancer on mammograms. Breast Cancer 1997;4:228–33.

(9) Warren Burhenne LJ, Wood SA, Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000;215: 554–62.

(10) Brem RF, Schoonjans JM. Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study. Clin Radiol 2001;56:150–4.

(11) Leichter I, Fields S, Nirel R, Bamberger P, Novak B, Lederman R, et al. Improved mammographic interpretation of masses using computer-aided diagnosis. Eur Radiol 2000;10:377–83.

(12) Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999;212:817–27.

(13) Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. Radiology 2001;220: 787–94.

(14) Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: lessons from recent experience. Acad Radiol 2002;9:1264–77.

(15) Brem RF. Enhancement of mammographic interpretation with computer-aided detection (CAD): a multi-institutional trial. Presented at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 25–30, 2001, Chicago, IL. Radiology 2001; 221(P):472.

(16) Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology 2001;220:781–6.

(17) Cupples TE. Impact of computer-aided detection (CAD) in a regional screening mammography program. Presented at the 87th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 25–30, 2001, Chicago, IL. Radiology 2001;221(P):520.

(18) Bandodkar P, Birdwell RL, Ikeda DM. Computer aided detection (CAD) with screening mammography in an academic institution: preliminary findings. Presented at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America, December 1–6, 2002, Chicago, IL. Radiology 2002;225(P):458.

(19) Young WW, Destounis SV, Bonaccio E, Zuley ML. Computer-aided detection in screening mammography: can it replace the second reader in an independent double read? Preliminary results of a prospective double blinded study. Presented at the 88th Scientific Assembly and Annual Meeting of the Radiological Society of North America, December 1–6, 2002, Chicago, IL. Radiology 2002;225(P):600.

(20) Food and Drug Administration. Quality Standards and Certification Requirements for Mammography Facilities (21 CFR Part 900). Federal Register 1993;58:67565.

(21) American College of Radiology (ACR). Breast imaging reporting and data system (BI-RADS). Reston (VA): American College of Radiology, 1998. Available at http://www.acr.org/departments/stand_accred/birads/content-s.html. [Last accessed: 12/9/03.]

(22) Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.

(23) Brown H, Prescott R. Applied mixed models in medicine. Chichester (NY): J. Wiley & Sons; 1999, pages 104–11

(24) Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR Am J Roentgenol 2001;177:543–9.

(25) Hoffmeister JW, Rogers SK, DeSimio MP, Brem RF. Determining efficacy of mammographic CAD systems. J Digit Imaging 2002;15 Suppl 1:198–200.

NOTES